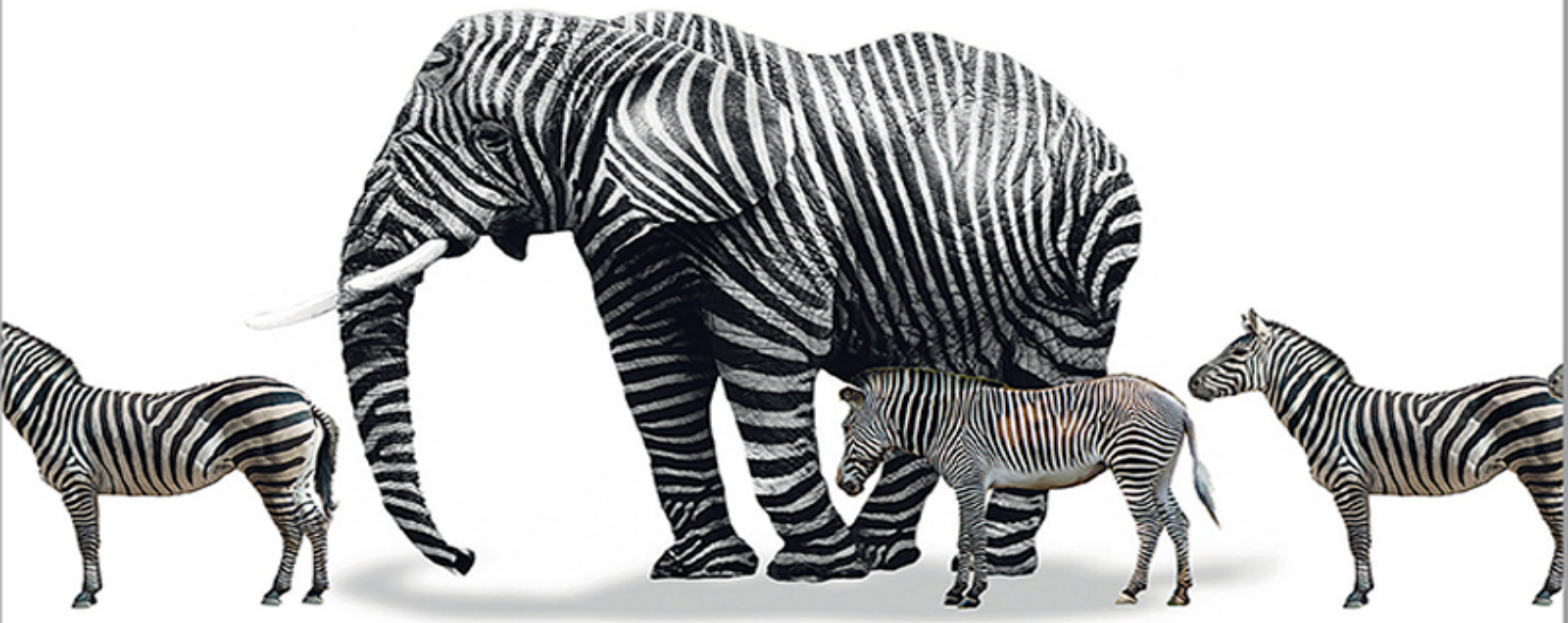
BESTSELLER „NEW YORK TIMESA”

WSZYSCY KŁAMIĄ

BIG DATA, NOWE DANE I WSZYSTKO, CO INTERNET MOŻE NAM POWIEDZIEĆ O TYM, KIM NAPRAWDĘ JESTEŚMY



SETH STEPHENS-DAVIDOWITZ

Seth Stephens-Davidowitz

WSZYSCY KŁAMIĄ

BIG DATA, NOWE DANE I WSZYSTKO, CO INTERNET MOŻE NAM POWIEDZIEĆ O TYM, KIM NAPRAWDĘ JESTEŚMY

PRZEDMOWA Steven Pinker

PRZEŁOŻYŁ Maciej Świerkocki

Wydawnictwo Literackie

Spis treści

Karta redakcyjna

Dedykacja

Przedmowa – Steven Pinker

Wstęp. Zarysy rewolucji

CZĘŚĆ PIERWSZA. ZBIORY DANYCH, DUŻE I MAŁE

1. Nasza omylna intuicja

CZĘŚĆ DRUGA. MOCE BIG DATA

2. Czy Freud miał rację?
3. Dane przeobrażone
    - Ciało jako zbiór danych
    - Słowa jako dane
    - Obrazy jako dane
4. Cyfrowe serum prawdy
    - Prawda o seksie
    - Prawda o nienawiści i uprzedzeniach
    - Prawda o Internecie
    - Prawda o przemocy wobec dzieci i aborcji
    - Prawda o waszych znajomych z Facebooka
    - Prawda o waszych klientach
    - Czy potrafimy się pogodzić z prawdą?

5. Zbliżenie

Co naprawdę dzieje się w amerykańskich hrabstwach, miastach i miasteczkach?

Jak spędzamy kolejne minuty i godziny swojego życia

Nasze sobowtóry

Opowieści z bazy danych

6. Świat to jedno wielkie laboratorium

Abecadło testów A/B

Okrutne – choć pouczające – eksperymenty natury

CZĘŚĆ TRZECIA. ZBIÓR BIG DATA: DLACZEGO NALEŻY OBCHODZIĆ SIĘ Z NIM OSTROŻNIE

7. Big data, big srata? Czego nie potrafi zbiór big data

Klątwa wymiarowości

Przesadny nacisk na to, co mierzalne

8. Więcej danych, więcej problemów? Czego nie powinniśmy robić

Zagrożenie ze strony wszechwładnych korporacji

Zagrożenie ze strony wszechwładnego państwa

Zakończenie. Ilu ludzi czyta książki do końca?

Podziękowania

Przypisy

Tytuł oryginału
Everybody Lies. Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are

Opieka redakcyjna: MACIEJ ZARYCH
Redakcja: KAMIL BOGUSIEWICZ
Korekta: URSZULA SROKOSZ-MARTIUK, ANETA TKACZYK, MARIA WOLAŃCZYK
Projekt okładki: LISA AMOROSO
Ilustracje na okładce: © Visuals Unlimited, Inc./Victor Habbick, © Shutterstock/Aaron Amat
Opracowanie graficzne wersji papierowej na podstawie oryginału: MAREK PAWŁOWSKI
Redakcja techniczna: ROBERT GĘBUŚ

Copyright © 2017 by Seth Stephens-Davidowitz.
All rights reserved
© Copyright for the Polish translation by Maciej Świerkocki
© Copyright for this edition by Wydawnictwo Literackie, 2019

Wydanie pierwsze

ISBN 978-83-08-07124-3

Wydawnictwo Literackie Sp. z o.o.
ul. Długa 1, 31-147 Kraków
tel. (+48 12) 619 27 70
fax. (+48 12) 430 00 96
bezpłatna linia telefoniczna: 800 42 10 40
e-mail: ksiegarnia@wydawnictwoliterackie.pl
Księgarnia internetowa: www.wydawnictwoliterackie.pl

Konwersja: eLitera s.c.

Mamie i Tacie

PRZEDMOWA

Odkąd filozofowie zaczęli spekulować na temat „cerebroskopu", mitycznego urządzenia wyświetlającego ludzkie myśli na ekranie monitora, przedstawiciele nauk społecznych szukają instrumentów, dzięki którym można by odkryć mechanizm natury ludzkiej. Przez długie lata mojej pracy psychologa eksperymentalnego wchodziły w modę i traciły na popularności różne metody badawcze, a ja próbowałem stosować je wszystkie – skalę ocen, czas reakcji, zwężenie źrenic czy neuroobrazowanie funkcjonalne. Miałem nawet chorych na epilepsję pacjentów z elektrodami, którzy w oczekiwaniu na kolejny atak chętnie poddawali się eksperymentom językowym.

Żadna z tych metod nie dostarcza jednak bezpośredniego wglądu w ludzki umysł, jesteśmy więc zmuszeni do wielkich kompromisów naukowych. Ludzkie myśli przyjmują postać skomplikowanych tez i twierdzeń; nie jest tak, jak chce Woody Allen – że przeczytawszy szybko *Wojnę i pokój*, myślimy potem tylko, że „To była książka o jakichś Rosjanach". A skomplikowane tezy i twierdzenia w swojej poplątanej, wielowymiarowej chwale sprawiają naukowcom trudności analityczne. Oczywiście gdy ludzie się przed nami otwierają, jesteśmy w stanie zarejestrować bogactwo ich strumienia świadomości, monologi nie stanowią jednak idealnej bazy danych, na których moglibyśmy sprawdzać swoje hipotezy. Z drugiej strony, jeśli skupimy się na elementach łatwo policzalnych, jak czas reakcji człowieka na słowa albo reakcja naszego naskórka na treści wizualne, możemy dokonywać rachunków statystycznych, ale redukujemy wtedy skomplikowaną strukturę poznawczą do pojedynczej liczby. Nawet najbardziej zaawansowane metody badań neuroobrazów mogą nam co najwyżej powiedzieć, jak myśl rozkłada się w przestrzeni 3-D, nie wyjaśnią jednak, z czego jest zbudowana.

Jak gdyby kompromis między łatwością obróbki materiału a jego złożonością nie był złem sam w sobie, uczeni zajmujący się naturą ludzką mają także kłopot z Prawem Małych Liczb. Taką nazwę Amos Tversky i Daniel Kahnemann nadali fałszywemu mniemaniu, że cechy populacji odbijają się w dowolnej, nawet najmniejszej próbce testowej. Wszyscy, w tym i uczeni biegli w matematyce, żywią żałośnie błędne poglądy co do tego, ile osób naprawdę trzeba przebadać, zanim będziemy mogli wyeliminować przypadkowe wnioski i domysły i uogólnić wyniki naszych eksperymentów w taki sposób, aby rozciągnąć je na wszystkich Amerykanów, nie wspominając już o całej ludzkości. A wartość naszych badań jest tym bardziej wątpliwa, jeśli uczestnicy biorą w nich udział dla osobistych korzyści, jak choćby nasi studenci trzeciego roku, którym wypłacamy drobne kwoty na piwo.

Ta książka opisuje zupełnie nowy sposób badania ludzkiego umysłu. Big data, wielki zbiór danych zbudowany na podstawie wyszukiwań internetowych i innych działań online, to wprawdzie nie cerebroskop, ale Seth Stephens-Davidowitz dowodzi, że takie zbiory danych pozwalają nam zajrzeć w głąb ludzkiej duszy głębiej niż kiedykolwiek. Pochyleni w samotności nad klawiaturą komputera ludzie przyznają się do najdziwniejszych rzeczy, czasami dlatego (na przykład na portalach randkowych albo szukając profesjonalnej porady), że chcą wywołać jakieś skutki w realnym życiu, a czasami przeciwnie, dlatego że ich nie chcą – zrzucają zatem niekiedy kamień z serca, przyznając się do pewnych pragnień czy obaw, bo wiedzą, że nikt w realnym świecie nie zareaguje na nie pełnym przerażenia zdumieniem albo w jakiś znacznie gorszy sposób. Czy jednak ktoś zareaguje, czy nie, internauci nie tylko stukają po prostu w klawisze lub przekręcają gałki, ale też wpisują w sieć tryliony najróżniejszych sekwencji znakowych, aby wyrazić za ich pośrednictwem swoje myśli w całym ich wybuchowym, kombinacyjnym ogromie. I jeszcze lepiej – pozostawiamy za sobą te cyfrowe ślady w takiej formie, że łatwo jest je gromadzić i analizować. Internauci wywodzą się z wszystkich grup społecznych. Mogą brać udział w niejawnych eksperymentach, w których wykorzystuje się różne bodźce i tabularyzuje je w czasie rzeczywistym. Ludzie dokonujący

wpisów w sieci skwapliwie dostarczają nam takich danych, i to we wprost gargantuicznych ilościach.

Książka *Wszyscy kłamią* jest czymś więcej niż tylko dowodem na słuszność pewnej koncepcji. Wnioski Stephensa-Davidowitza wielokrotnie przewracały do góry nogami mniemania, jakie miałem na temat mojego kraju i gatunku *homo sapiens*. Skąd wzięło się nieoczekiwane poparcie dla Donalda Trumpa? Czy kiedy w 1976 roku Ann Landers zapytała swoich czytelników, czy żałują tego, że mają dzieci, i ze zdumieniem stwierdziła, że większość z nich odpowiada twierdząco, to została wprowadzona w błąd, bo jej sondaż objął niereprezentatywną próbkę respondentów, których sama wybrała? Czy to Internet ponosi winę za powstanie „bańki filtrującej", jak redundantnie bywa nazywany sieciowy kryzys końca drugiej dekady XXI wieku? Co powoduje przestępstwa nienawiści? Czy ludzie szukają dowcipów w Internecie, by poprawić sobie nastrój? Lubię myśleć, że nic nie jest w stanie mnie zaszokować, lecz informacje o tym, co Internet mówi nam na temat ludzkiej seksualności, naprawdę mną wstrząsnęły – jak na przykład wiadomość, że określona liczba kobiet co miesiąc wyszukuje zdjęcia oraz filmy przedstawiające „ruchanie wypchanych zwierząt". Tego faktu nie mogłyby ujawnić żadne eksperymenty oparte na czasie reakcji, zwężeniu źrenic czy funkcjonalnym neuroobrazowaniu.

Wszyscy kłamią przypadnie do gustu każdemu czytelnikowi. Stephens-Davidowitz z niezłomną dociekliwością i czarującą inteligencją wskazuje nową ścieżkę rozwoju dla nauk społecznych w XXI stuleciu. Po co nam cerebroskop, skoro otworzyliśmy tak nieskończenie fascynujące okno, przez które dobrze widać nasze obsesje?

Steven Pinker, 2017

WSTĘP
ZARYSY REWOLUCJI

Wszyscy mówili, że na pewno przegra.

W czasie prawyborów w Partii Republikańskiej w 2016 roku specjaliści od ankiet i sondaży doszli do przekonania, że Donald Trump nie ma najmniejszych szans na zwycięstwo. Przecież obraził dużo różnych mniejszości. A z badań opinii publicznej i wypowiedzi ich interpretatorów wynikało, że tylko bardzo niewielu Amerykanów toleruje takie oburzające wystąpienia.

Większość ekspertów sądziła też, że Trump przegra w wyborach powszechnych. Zbyt wielu potencjalnych wyborców dawało do zrozumienia, że nie podobają im się jego postawa i poglądy.

Pewne oznaki świadczące o tym, że Donald Trump mimo wszystko może zwyciężyć zarówno w prawyborach, jak i w wyścigu o fotel prezydenta, jednak się pojawiły – w Internecie.

Jestem specjalistą od analizy danych internetowych. Codziennie tropię różne cyfrowe znaki, jakie ludzka obecność zostawia za sobą w sieci. Na podstawie śladów po klawiszach, które wciskamy na klawiaturze komputera, albo kliknięciach myszką, próbuję odgadnąć, czego naprawdę pragniemy, jak rzeczywiście postąpimy w przyszłości i kim faktycznie jesteśmy. Pozwólcie teraz, że wytłumaczę wam, co sprawiło, że wstąpiłem na tę dosyć niezwykłą drogę życiową.

Moja historia zaczyna się wraz z amerykańskimi wyborami prezydenckimi w 2008 roku – dzisiaj wydaje się, że było to już całe wieki temu – i łączy z pewnym pytaniem, które od lat pojawia się

w rozważaniach przedstawicieli nauk społecznych: jak ważną rolę odgrywają w Stanach Zjednoczonych uprzedzenia rasowe?

Barack Obama jako pierwszy Afroamerykanin w historii startował w wyborach z ramienia wielkiej partii politycznej, nominowany do walki o fotel prezydenta. Wygrał – i to dość łatwo, a badania opinii publicznej zdawały się sugerować, że kolor jego skóry nie miał większego znaczenia w głosowaniu[1]. Na przykład Instytut Gallupa przeprowadził wiele sondaży przed pierwszą elekcją Obamy oraz po niej. Wnioski? Większości głosujących nie obchodziło to, że kandydat demokratów jest czarny. Niedługo po wyborach dwaj znani profesorowie z University of California w Berkeley przekopali się przez inne dane, oparte na badaniach opinii publicznej, stosując bardziej wyrafinowane metody analityczne[2], doszli jednak do takiej samej konkluzji.

Dlatego za czasów prezydentury Obamy opinia ta przeważała, zarówno w większości mediów, jak i ośrodków akademickich. Metody, do których środki przekazu i socjologowie odwoływali się wtedy już od ponad osiemdziesięciu lat, próbując zrozumieć rzeczywistość społeczną, dostarczały nam dowodów na to, że przy podejmowaniu decyzji o tym, czy Barack Obama ma zostać prezydentem, znaczna większość obywateli Stanów Zjednoczonych nie brała pod uwagę jego etnicznego pochodzenia.

Wyglądało na to, że w USA, przez długie lata plamionych niewolnictwem i dyskryminującymi czarnych ustawami, przestano w końcu oceniać ludzi po kolorze skóry, co z kolei zdawało się sugerować, że amerykański rasizm dogorywa. Znaleźli się nawet mędrcy, którzy ogłosili, że żyjemy już w społeczeństwie postrasowym[3].

W 2012 roku, gdy Obama zwyciężył po raz drugi, byłem zagubionym życiowo doktorantem ekonomii – wypalonym, jeśli chodzi o moją pierwszą specjalność naukową, choć przy tym pewnym siebie, a nawet pyszałkowatym, sądziłem bowiem, że całkiem już nieźle rozumiem, na czym polega ten świat oraz co ludzie XXI wieku myślą i do czego przywiązują wagę. Jeżeli chodzi o uprzedzenia rasowe, to bazując na swoich lekturach psychologicznych i politologicznych, mniemałem, że

do rasizmu otwarcie przyznaje się zaledwie znikomy odsetek mieszkańców Stanów Zjednoczonych, w większości konserwatywnych republikanów, mieszkających na ogół na dalekim południu kraju.

Potem odkryłem jednak Google Trends.

Wyszukiwarka ta, wypuszczona po raz pierwszy na rynek bez większych fanfar w 2009 roku, informuje swoich użytkowników, jak często internauci szukali w sieci danego słowa czy frazy w różnych miejscach i czasie. Reklamowano ją właściwie jako zabawkę, na przykład umożliwiającą grupie znajomych dyskusję o tym, który celebryta cieszy się największą popularnością albo co nieoczekiwanie zrobiło się „trendy". Jej najwcześniejsze wersje opatrywano humorystycznym zastrzeżeniem, że na podstawie takich danych źródłowych „nikt nie zechce napisać doktoratu", co oczywiście natychmiast mnie skłoniło, aby napisać doktorat, korzystając właśnie z nich[1*].

Informacji pochodzących z wyszukiwarki Google nie uważano w tamtym czasie za odpowiednie źródło dla „poważnych" badań akademickich. W przeciwieństwie do ankiet i sondaży, Google'a nie stworzono z myślą o tym, by pomóc nam zrozumieć ludzką psychikę, tylko raczej po to, żebyśmy mogli dowiedzieć się czegoś o świecie, wyszukiwarki nie traktowano zatem jako narzędzia, dzięki któremu uczeni mogliby lepiej poznać naturę ludzką. Okazało się jednak, że ślady, które po sobie pozostawiamy podczas wyszukiwania w sieci, są wielce pouczające.

Inaczej mówiąc, szukanie informacji samo jest informacją. To, kiedy i gdzie ludzie poszukują różnych rzeczy, faktów, cytatów, dowcipów, miejsc, osób lub pomocy, może, jak się okazuje, powiedzieć nam znacznie więcej, niż większość z nas przypuszcza, o tym, co naprawdę myślimy i robimy oraz czego naprawdę pragniemy i się boimy. Wniosek taki jest tym bardziej uprawniony, że ludzie czasem nie tyle zadają Google'owi pytania, ile mu się zwierzają: „Nie cierpię swojego szefa". „Jestem pijany". „Tata mnie uderzył".

Codzienna czynność polegająca na tym, że wklepujemy jakieś słowo czy frazę w niewielki prostokątny pasek wyszukiwarki, pozostawia po sobie mikroskopijny ślad prawdy, który pomnożony przez wiele

milionów, doprowadza nas w końcu do pewnych istotnych faktów. Pierwszym słowem, które wpisałem w Google Trends, był „Bóg" – i dowiedziałem się, że stany, gdzie najczęściej pojawiają się zawierające je wyszukiwania, to Alabama, Missisipi i Arkansas – czyli stany należące do tak zwanego Pasa Biblijnego. I że wyszukiwania te prowadzone są najczęściej w niedziele. Żadna z tych informacji nie była zaskakująca, lecz zaintrygowało mnie to, że z danych o wyszukiwaniu może wynikać tak konkretna wiedza. Następnie wpisałem nazwę drużyny koszykarskiej „Knicks" i okazało się, że pojawia się ona najczęściej w wyszukiwaniach prowadzonych w Nowym Jorku. Tego też nietrudno byłoby się domyślić. W końcu wpisałem w Internet swoje nazwisko. „Przepraszamy – poinformowała mnie Google Trends. – Brak mocy obliczeniowej, żeby wyświetlić rezultaty". Dzięki temu zdałem sobie sprawę, że Google Trends może udzielić odpowiedzi tylko wtedy, kiedy poszukiwania w sieci na taki czy inny temat prowadzi wiele osób.

Siła odpowiedzi Google'a nie polega jednak na tym, że można się dowiedzieć, że Bóg cieszy się zainteresowaniem na Południu, zespół Knicksów w Nowym Jorku, a mną nie interesuje się zupełnie nikt i nigdzie. O tym mogłaby nas przekonać każda ankieta. Siłą Google'a jest to, że ludzie mówią tej gigantycznej wyszukiwarce rzeczy, jakich nie powiedzieliby nikomu innemu.

Weźmy na przykład seks (temat, któremu przyjrzę się znacznie bliżej w dalszej części książki). W kwestii naszego życia seksualnego nie można ufać sondażom i ankietom. Przeanalizowałem na przykład informacje z badań General Social Survey[4], uważanych za jedno z najbardziej wpływowych i autorytatywnych źródeł informacji o zachowaniach i zwyczajach Amerykanów i Amerykanek. Zgodnie z wynikami GSS – jeśli chodzi o seks par hetero – Amerykanki twierdzą, że odbywają średnio pięćdziesiąt pięć stosunków seksualnych rocznie, w szesnastu procentach przypadków z użyciem kondomów. Daje to miliard sto tysięcy zużytych prezerwatyw na rok. Heteroseksualni mężczyźni utrzymują jednak, że zużywają rocznie miliard sześćset tysięcy kondomów – a przecież liczby podawane przez kobiety i mężczyzn powinny być w tym kontekście jednakowe niejako z definicji. Kto zatem mówi prawdę, mężczyźni czy kobiety?

Okazuje się, że ani jedna, ani druga strona. Zgodnie z analizą globalnej firmy Nielsen, zajmującej się zbieraniem informacji i badaniami metrycznymi zachowań konsumenckich, w Stanach Zjednoczonych sprzedaje się mniej niż sześćset milionów kondomów na rok[5]. Kłamiemy zatem wszyscy, różnica polega tylko na tym, w jakim stopniu mijamy się z prawdą.

A nasze kłamstwa mają daleki zasięg. Na przykład panowie, którzy nigdy nie byli żonaci, twierdzą, że zużywają średnio dwadzieścia dziewięć kondomów rocznie, co oznacza, że ich suma przewyższyłaby liczbę wszystkich prezerwatyw sprzedanych w USA zarówno mężczyznom pozostającym w związkach małżeńskich, jak i kawalerom. Osoby żonate i zamężne też chyba przesadzają, opowiadając o tym, jak często uprawiają seks – bo żonaci mężczyźni poniżej sześćdziesiątego piątego roku życia twierdzą, że kochają się przeciętnie raz w tygodniu, a zaledwie jeden procent z nich przyznaje, że w ostatnim roku nie uprawiał seksu ani razu. Mężatki, jak wynikałoby z badań, kochają się nieco rzadziej, ale tylko trochę.

Tymczasem Google przedstawia daleko mniej ożywiony – i w moim mniemaniu o wiele prawdziwszy – obraz naszego małżeńskiego życia erotycznego. Według wyszukiwarki najczęściej skarżymy się w małżeństwie na brak seksu. Wpisy takie jak „brak seksu małżeńskiego" są trzy i pół razy częstsze niż „nieszczęśliwe małżeństwo" i osiem razy częstsze niż „małżeństwo bez miłości". Zresztą nawet pary pozostające w związkach partnerskich nierzadko narzekają na brak seksu. Częstotliwość wyszukiwania fraz typu „związek bez seksu" ustępuje tylko wyszukiwaniu „przemocy w związku" (muszę tutaj dodać, że wszystkie te informacje podawane są anonimowo, Google nie ujawnia bowiem oczywiście żadnych danych dotyczących wyszukiwań prowadzonych przez osoby przedstawiane z imienia i nazwiska).

Poza tym wpisy w Google'u pokazywały zupełnie inny obraz Ameryki niż postrasowa utopia, jaką odmalowują badania ankietowe. Pamiętam, kiedy po raz pierwszy wpisałem w Google Trends słowo „czarnuch". Nazwijcie mnie naiwniakiem, lecz biorąc pod uwagę wysoką toksyczność tego określenia, naprawdę oczekiwałem, że nie będzie

używane w Internecie zbyt często. Grubo się myliłem. Bo w Stanach Zjednoczonych „czarnuch" – albo jego liczba mnoga, „czarnuchy" – pojawiało się w wyszukiwarce mniej więcej równie często jak „migrena" („migreny"), „ekonomista" czy „Lakersi". Zacząłem się zastanawiać, czy na takie wyniki nie mają wpływu teksty piosenek raperskich. Ale nie. A to dlatego, że słowo, którego używają raperzy w swoich utworach, pisze się inaczej: nie „nigger" („niggers"), tylko „nigga(s)". Co zatem kierowało Amerykanami wstukującymi „czarnuchów" w wyszukiwarkę? Często szukali dowcipów o Afroamerykanach – dwadzieścia procent wyszukiwań ze słowem „czarnuch" obejmowało również słowo „dowcipy". Inne często spotykane tego typu hasła w Google'u to „głupie czarnuchy" i „nienawidzę czarnuchów".

Podobnych wpisów co roku pojawiały się miliony. Duża liczba Amerykanów przeprowadzała w swoim zaciszu domowym szokująco rasistowskie poszukiwania w sieci. Im bliżej badałem tę sprawę, tym więcej znajdowałem niepokojących danych.

Pierwszego wieczoru po wyborze Obamy na prezydenta, gdy większość komentarzy skupiała się na komplementowaniu go i podkreślaniu historycznego znaczenia jego elekcji, w Google'u mniej więcej jeden wpis na sto, które zawierały słowo „Obama", zawierał również skrótowiec „kkk" albo określenie „czarnuch(y)". Może nie wydaje się, że to tak wiele, pomyślmy jednak o tysiącach nierasistowskich powodów, dla jakich można by wygooglować tego młodego czarnoskórego outsidera z czarującą rodziną, który właśnie dostał pracę na stanowisku najpotężniejszego człowieka na świecie. W dniu wyborów liczba wyszukiwań i wejść na Stormfront[6] – zaskakująco popularny w USA portal białych nacjonalistów – była dziesięciokrotnie wyższa niż zazwyczaj. W niektórych stanach znalazłem więcej wpisów w rodzaju „prezydent czarnuch" niż „pierwszy czarny prezydent"[7].

Mieliśmy więc do czynienia z ciemnotą i nienawiścią, nieobecnymi w tradycyjnych środkach przekazu, ale na porządku dziennym w Internecie.

Wyszukiwania tej treści trudno jest pogodzić z obrazem społeczeństwa, w którym rasizm odgrywa rzekomo niewielką rolę.

W 2012 roku Donald J. Trump był mi znany głównie jako biznesmen i uczestnik reality shows. Jak większość z nas, nie miałem bladego pojęcia, że cztery lata później stanie się poważnym pretendentem do fotela prezydenckiego. Ale internetowe wpisy o tak paskudnym rasistowskim wydźwięku łatwo jest chyba zrozumieć na tle sukcesu kandydata, który w swoich atakach na imigrantów, gniewnych wystąpieniach i resentymentach grał często na najniższych instynktach ludzkich.

Wyszukiwania w Google'u uświadomiły nam też, że wiele naszych wyobrażeń na temat geografii uprzedzeń rasowych jest mylnych. Badania ankietowe i powszechne przekonania umieszczały centrum współczesnego rasizmu głównie na południu USA i wśród republikanów. Ale do miejsc, gdzie rasistowskie wpisy pojawiały się najczęściej, należy zaliczyć północną część stanu Nowy Jork, zachodnią Pensylwanię, wschodnie Ohio, przemysłowe Michigan i rolnicze Illinois, choć także Wirginię Zachodnią, południową Luizjanę oraz Missisipi. Prawdziwa linia podziału, jak sugerowały dane z Google'a, nie przebiegała na granicy między południem i północą USA, tylko pomiędzy wschodem i zachodem. Na zachód od Missisipi nie jest już tak źle. Poza tym rasizm bynajmniej nie ograniczał się do republikanów. Mówiąc wprost, rasistowskie wyszukiwania wcale nie występowały częściej w środowisku republikańskim niż wśród demokratów. Innymi słowy, wyszukiwarka Google pomogła nam nakreślić nową mapę rasizmu w Stanach Zjednoczonych, wyglądającą zupełnie inaczej, niż można by przypuszczać. Republikanie z Południa prędzej przyznają się publicznie do rasizmu, lecz wielu demokratów z Północy w głębi duszy podziela ich poglądy.

Mapa, którą stworzyłem dzięki wpisom w Google'u, miała się okazać niezwykle ważna dla wyjaśnienia politycznego sukcesu Trumpa cztery lata później.

W 2012 roku wykorzystałem ją po to, aby ponownie ocenić rolę, jaką odegrało w wyborach pochodzenie etniczne Baracka Obamy. Dane okazały się jednoznaczne. W tych regionach kraju, gdzie wyszukiwań rasistowskich było wiele, Obama uzyskał znacząco gorsze wyniki niż

cztery lata wcześniej John Kerry, biały kandydat demokratów na prezydenta. Nie dało się wyjaśnić tej różnicy innymi czynnikami, takimi jak poziom wykształcenia, wiek wyborców, ich religijność czy łatwość dostępu do broni w danym stanie. Ale to, że spośród kandydatów demokratycznych jedynie Barack Obama osiągnie słaby wynik, można było przewidzieć wyłącznie na podstawie internetowych wpisów o treściach rasistowskich.

A wpisy te miały niebagatelne znaczenie. Na skutek postaw jawnie rasistowskich Obama stracił mniej więcej cztery punkty procentowe w skali kraju, czyli znacznie więcej, niż mogliśmy się spodziewać według sondaży. Oczywiście został wybrany na prezydenta dwukrotnie, w czym dopomogły mu pewne niezwykle sprzyjające Partii Demokratycznej okoliczności, musiał jednak pokonać znacznie większe przeszkody, niż spodziewali się ludzie opierający swoje przewidywania na tradycyjnych źródłach informacji – czyli właściwie wszyscy. Rasistów znalazło się zaś potem dosyć, by w roku wyborczym już nie tak pomyślnym dla demokratów pomóc wygrać prawybory ich przeciwnikom albo przechylić szalę na niekorzyść Partii Demokratycznej w wyborach powszechnych.

Moja analiza została początkowo odrzucona przez aż pięć pism akademickich[8]. Wielu recenzentów, jeśli czytelnik wybaczy mi pewną zgryźliwość, twierdziło, że nie sposób uwierzyć, aby tak wielu Amerykanów hołdowało takim zajadłym poglądom rasistowskim. Po prostu ten obraz nie pasował do oficjalnych wypowiedzi naszych obywateli. Poza tym wyszukiwarka Google'a wielu uczonym wydawała się jeszcze wtedy mocno kuriozalną bazą danych.

Dziś, po zaprzysiężeniu prezydenta Donalda J. Trumpa, moje ustalenia wyglądają chyba o wiele wiarygodniej.

Im dłużej prowadziłem swoje badania, tym bardziej się przekonywałem, że Google zawiera wiele informacji, których nie uwzględniają sondaże, a które mogą nam pomóc w zrozumieniu mechanizmu wyborów prezydenckich – a także wielu, bardzo wielu innych zjawisk.

Można na przykład znaleźć w Google'u informacje o tym, kto ostatecznie pójdzie do urn wyborczych. Ponad połowa obywateli, którzy nie głosują, deklaruje w sondażach tuż przed wyborami, że zamierza głosować, co wpływa na nasze przewidywania frekwencji; tymczasem wyszukiwania o treści „jak głosować" albo „gdzie głosować", prowadzone w Internecie nawet na wiele tygodni przed wyborami, mogą nam pomóc precyzyjnie ustalić, w których częściach kraju frekwencja będzie najwyższa.

W sieci możemy znaleźć nawet informację, na kogo ci wyborcy oddadzą swój głos. Ja i Stuart Gabriel, profesor bankowości i finansów z University of California w Los Angeles, odkryliśmy zaskakującą wskazówkę, pozwalającą odgadnąć, jak będą głosować Amerykanie. Otóż znaczny odsetek wyszukiwań związanych z wyborami zawiera pytania z nazwiskami obu kandydatów. W czasie wyborów w 2016 roku, kiedy kandydatami byli Trump i Hillary Clinton, niektórzy wpisywali w wyszukiwarkę hasło „sondaże Trump Clinton", inni zaś szukali najciekawszych momentów z „debaty Clinton Trump". Okazało się, że dwanaście procent wyszukiwań z nazwiskiem „Trump" zawierało także nazwisko „Clinton", a ponad jedna czwarta z hasłem „Clinton" zawierała też słowo „Trump".

Gabriel i ja ustaliliśmy, że te pozornie neutralne wyszukiwania mogą nam pomóc wskazać kandydata, którego popierają internauci.

W jaki sposób? Chodzi o kolejność, w jakiej wymieniani są rywale. Nasze badania dowodzą, że ludzie mają skłonność na pierwszym miejscu podawać nazwisko tego kandydata, którego popierają.

W czasie trzech ostatnich wyborów prezydenckich kandydat, który w wyszukiwaniach występował częściej jako pierwszy, dostawał więcej głosów. A co jest jeszcze ciekawsze, porządek, w jakim wyszukiwano pretendentów do Białego Domu, pozwalał przewidzieć wyniki głosowania w poszczególnych stanach.

Okazało się to możliwe dlatego, że kolejność, w której wymieniamy rywali, również wydaje się zawierać pewne informacje, niewynikające z badania opinii publicznej. W 2012 roku, kiedy wyborcy decydowali pomiędzy Obamą i republikaninem Mittem Romneyem, Nate Silver, wirtuoz statystyki oraz dziennikarstwa, trafnie przewidział wyniki

głosowania w pięćdziesięciu stanach. My jednak ustaliliśmy, że tam, gdzie w wyszukiwarce na pierwszym miejscu częściej figurował Romney, uzyskiwał on lepszy wynik niż przewidywany przez Silvera, a w stanach, gdzie internauci częściej wpisywali na pierwszym miejscu Obamę, jego wyniki również były lepsze od tych, które prognozował Silver.

Wskazywałoby to, że sondaże mylą się dlatego, że wyborcy okłamują samych siebie albo nie bardzo mają ochotę zdradzać ankieterom swoje rzeczywiste preferencje. Jeśli w roku 2012 twierdzili, że są niezdecydowani, lecz konsekwentnie wpisywali w wyszukiwarkę hasła w rodzaju „sondaże Romney Obama", „debata Romney Obama" czy „wybory Romney Obama", to najprawdopodobniej po cichu przez cały czas zamierzali głosować na Romneya.

Czy zatem wyszukiwarka Google niejako przewidziała zwycięstwo Trumpa? No cóż, czeka nas w tej kwestii jeszcze sporo pracy i będzie musiało dołączyć do mnie znacznie więcej badaczy, zanim dowiemy się, jak najlepiej korzystać z materiałów z Google'a, aby przewidywać wyniki głosowania. Analiza danych to nowa gałąź nauki, a my dysponujemy jedynie informacjami z kilku ostatnich elekcji. I z całą pewnością nie twierdzę, że dotarliśmy albo że w ogóle kiedykolwiek dotrzemy do punktu, w którym sondaże opinii publicznej, rozumiane jako narzędzie pomagające przewidzieć wynik wyborów, będzie można po prostu wyrzucić na śmietnik.

Nie zmienia to jednak faktu, że w Internecie dało się zauważyć wiele oznak świadczących o tym, że Trump uzyska lepszy rezultat, niż wskazywały badania opinii publicznej.

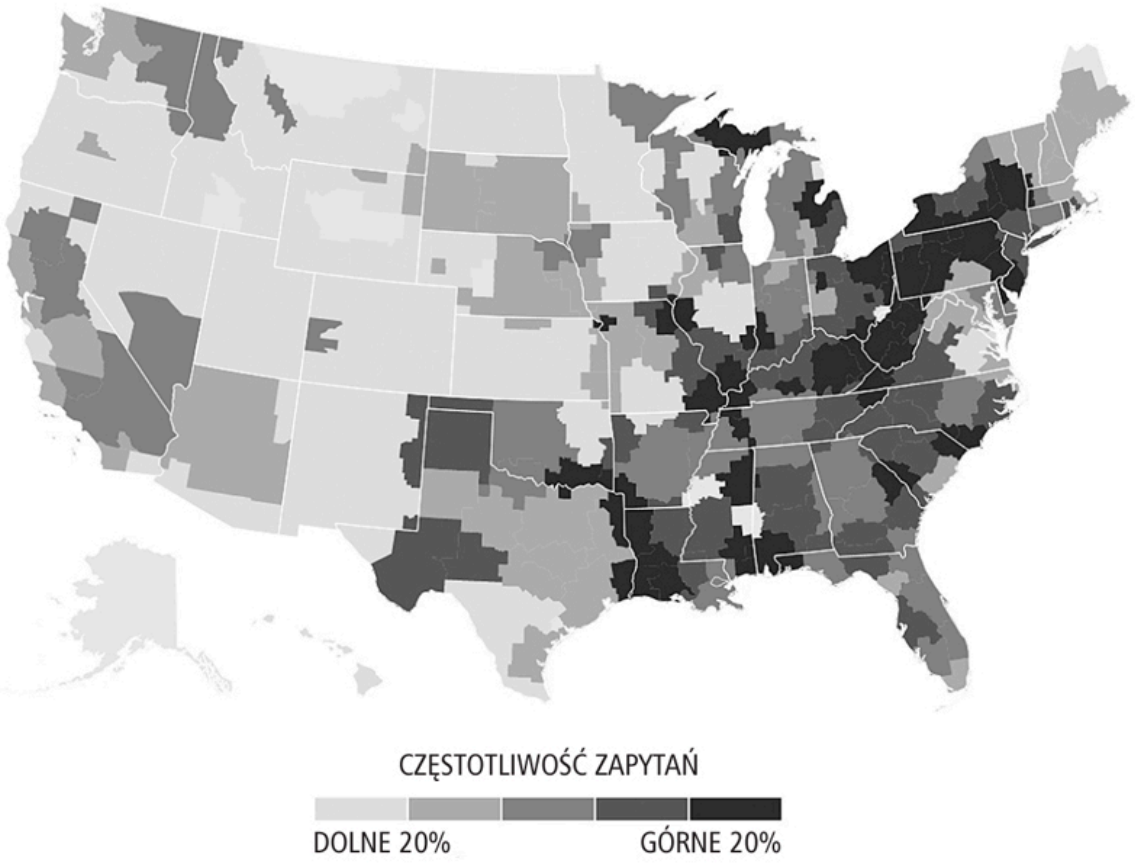
Później, już w czasie wyborów powszechnych, wystąpiły kolejne sygnały, że elektorat może sprzyjać Trumpowi. Czarni Amerykanie wciąż podawali w sondażach, że pójdą licznie do urn, żeby wystąpić przeciwko niemu. Ale liczba wyszukiwań na temat wyborów w rejonach gęsto zamieszkanych przez czarnych była niewielka, a w dniu elekcji Hillary Clinton boleśnie odczuła niską frekwencję Afroamerykanów.

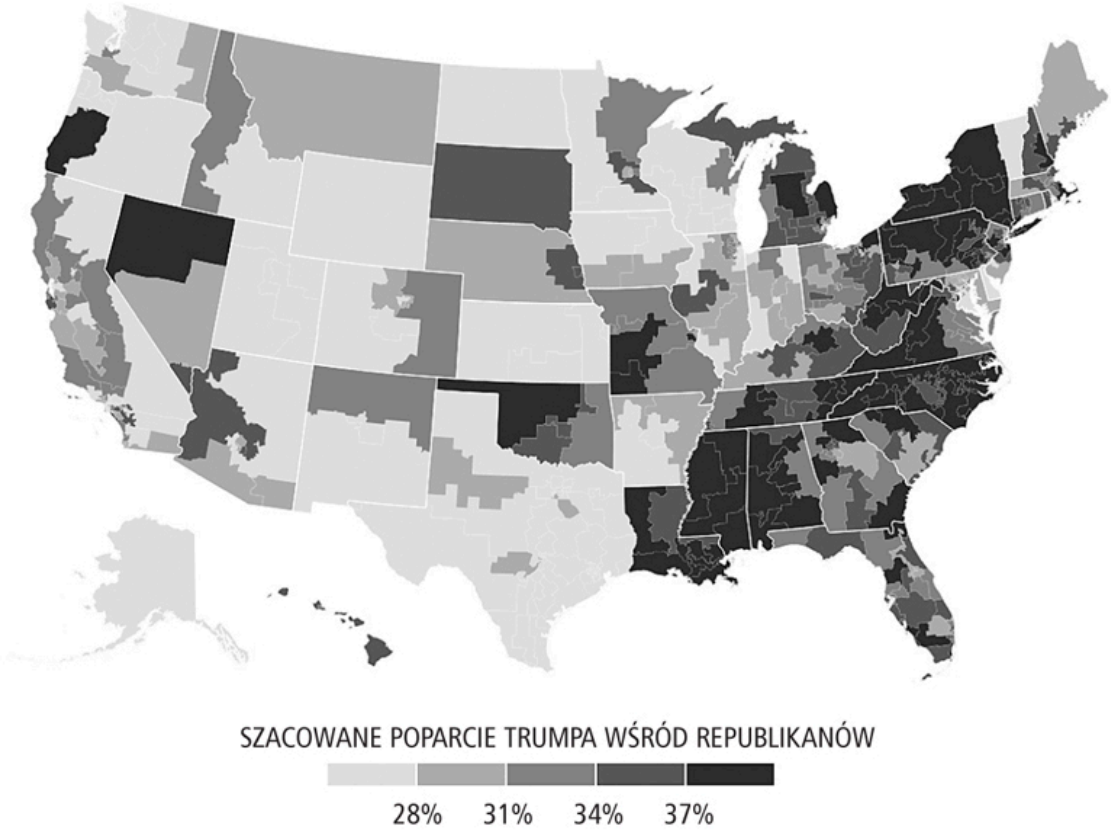
Pojawiły się też pewne oznaki, że rzekomo niezdecydowani wyborcy wolą jednak Trumpa. Gabrielowi i mnie udało się bowiem wykazać, że w kluczowych stanach Środkowego Zachodu, w których spodziewano

się zwycięstwa Clinton, internauci znacznie częściej wpisywali w wyszukiwarkę hasło „Trump Clinton" niż „Clinton Trump". Można wręcz powiedzieć, że Trumpowi udało się ostatecznie wygrać właśnie dlatego, że uzyskał w tych stanach wyniki wyraźnie lepsze od prognozowanych.

Upierałbym się jednak, że główną wskazówką dowodzącą możliwości Trumpa – na początek w prawyborach – był ów utajony rasizm, który odkryłem, badając elekcje wygrane przez Obamę. Internet ujawnił ciemnotę i nienawiść znacznej liczby Amerykanów, które przez wiele lat uchodziły uwadze ekspertów. Dane internetowe pokazały, że jesteśmy zupełnie innym społeczeństwem niż to, w którym żyjemy zdaniem akademików i dziennikarzy, opierających swoje opinie na badaniach ankietowych. Pokazały obrzydliwą, przerażającą i powszechną nienawiść, którą nowy kandydat prezydencki uzewnętrzniał niejako w imieniu swoich zwolenników.

Wskaźnik rasistowskich zapytań





Ludzie często kłamią – okłamują siebie oraz innych. W 2008 roku Amerykanie w badaniach opinii publicznej twierdzili, że kwestie rasowe nic ich już nie obchodzą, a osiem lat później wybrali na prezydenta Donalda J. Trumpa, człowieka, który rozpowszechniał na Twitterze fałszywą informację, jakoby to czarni byli odpowiedzialni za większość zabójstw białych Amerykanów, bronił swoich popleczników, którzy poturbowali demonstranta z organizacji Black Lives Matter na jednym z wieców wyborczych, i długo się wahał, czy powinien zrezygnować z poparcia ze strony byłego przywódcy Ku-Klux-Klanu. Utajony rasizm, który zaszkodził Barackowi Obamie, dopomógł Trumpowi.

Na początku prawyborów Nate Silver wygłosił swoją słynną już opinię, że Trump właściwie nie ma szans na zwycięstwo. W miarę upływu czasu okazywało się jednak coraz wyraźniej, że kandydat republikanów cieszy się szerokim poparciem, Silver postanowił więc przyjrzeć się danym analitycznym i zobaczyć, czy uda mu się na tej podstawie zrozumieć, co się dzieje. Jak to możliwe, że Trump osiąga takie dobre wyniki?

Silver zauważył, że regiony, w których Trumpowi szło najlepiej, tworzą dziwnie układającą się mapę, przyszły prezydent radził sobie bowiem świetnie w wielu miejscach na Północnym Wschodzie i uprzemysłowionym Środkowym Zachodzie, nie tylko na Południu, znacznie słabiej wypadł natomiast w stanach zachodnich. Nate zaczął więc szukać odpowiednich zmiennych, żeby wyjaśnić ten układ. Czy decydowało o nim bezrobocie? Religia? Dostęp do broni palnej? Liczba imigrantów? Niechęć do Obamy?

Doszedł do tego samego wniosku, co ja cztery lata wcześniej[9]. Trump mógł liczyć na największe poparcie tam, gdzie w wyszukiwarkach najczęściej pojawiało się słowo „czarnuch".

W ciągu ostatnich czterech lat analizowałem dane z wyszukiwarki właściwie codziennie, a przez pewien czas nawet jako analityk w Google'u, bo firma wynajęła mnie do pracy, dowiedziawszy się o moich internetowych badaniach nad rasizmem. Nadal analizuję dane jako felietonista i dziennikarz, piszący o Internecie dla „New York Timesa". I nieustannie odkrywałem coś nowego. Interesowały mnie choroby umysłowe, ludzka seksualność, molestowanie dzieci, aborcja, reklama, religia, zdrowie. To żywotne zagadnienia, a nieistniejące jeszcze dwadzieścia lat temu materiały źródłowe rzucały na nie zaskakująco nowe światło. Ekonomiści oraz inni uczeni, zajmujący się naukami społecznymi, nieustannie poszukują nowych źródeł informacji, powiem więc bez ogródek: jestem dziś przekonany, że wyszukiwania w Google'u to najważniejszy w dziejach ludzkości zbiór danych dotyczących psychiki człowieka.

Ale nie jest to jedyne narzędzie, jakie oferuje nam Internet, abyśmy mogli lepiej zrozumieć świat. Szybko się zorientowałem, że istnieją jeszcze inne cyfrowe kopalnie złota. Zapisałem sobie na twardym dysku całą Wikipedię, przeglądałem profile na Facebooku i „przeleciałem" przez Stormfront. PornHub, jeden z największych portali pornograficznych, udostępnił mi kompletne dane na temat treści wyszukiwanych i oglądanych w sieci przez anonimowych użytkowników z całego świata. Inaczej mówiąc, zanurzyłem się bardzo głęboko w czymś, co nazywamy dzisiaj wielkim zbiorem danych – big

data. Poza tym przeprowadziłem dziesiątki rozmów z innymi osobami – uczonymi, dziennikarzami i menedżerami – które również zajmują się tymi nowymi aspektami rzeczywistości, i wiele wyników ich badań przedstawiam w niniejszej książce.

Najpierw jednak muszę się do czegoś przyznać: nie podam dokładnej definicji big data. Dlaczego? Ponieważ jest to pojęcie nieprecyzyjne z natury. Co znaczy „wielki"? Czy osiemnaście tysięcy czterysta sześćdziesiąt dwie obserwacje to jeszcze „small data", a osiemnaście tysięcy czterysta sześćdziesiąt trzy to już jest duży zbiór danych? Preferuję podejście całościowe: wprawdzie większość informacji, którymi się zajmuję, pochodzi z Internetu, będę też jednak omawiał inne źródła. Żyjemy w epoce zarówno ilościowej, jak i jakościowej eksplozji danych z wszystkich możliwych dziedzin. Wiele z tych nowych informacji napływa z Google'a i mediów społecznościowych, choć w pewnej mierze są to zapisane dziś cyfrowo stare treści, tkwiące gdzieś do tej pory w różnych szafkach i teczkach. Część materiałów źródłowych pochodzi z mnożących się wciąż ośrodków badań rynkowych i opinii publicznej. A jeszcze inne badania omawiane w mojej książce w ogóle nie korzystają z wielkich zbiorów danych, lecz po prostu stosują nowe, kreatywne podejście do nich – co w epoce nadmiaru informacyjnego jest niesłychanie istotne.

Dlaczego więc big data to tak potężny zbiór? Pomyślcie o wszystkich informacjach, które rozsiewamy w sieci każdego dnia – a potrafimy nawet określić ich liczbę. Otóż dziś, w drugiej dekadzie XXI wieku, ludzie produkują średnio dwa i pół miliona trylionów bajtów informacji dziennie[10].

I te bajty stanowią dla nas wskazówki.

Pewna kobieta nudzi się w czwartkowe popołudnie. Googluje w poszukiwaniu nowych „śmiesznych nieświńskich dowcipów". Sprawdza swoją skrzynkę w poczcie elektronicznej. Zakłada konto na Twitterze. Googluje w poszukiwaniu „dowcipów o czarnuchach".

Pewnego mężczyznę ogarnia smutek. Googluje w poszukiwaniu „symptomów depresji" i „opowieści o depresji". Układa elektronicznego pasjansa.

Inna kobieta dostaje na Facebooku wiadomość, że jej przyjaciółka się zaręczyła. Adresatka, która jest singielką, blokuje dalsze wiadomości od niej.

Inny mężczyzna robi sobie przerwę w googlowaniu informacji o zespołach hokejowych w NHL i muzyce rap, aby wstukać w wyszukiwarkę pytanie: „Czy to normalne, że mi się śni, że całuję się z mężczyznami?".

Jakaś kobieta klika w nagłówek w BuzzFeed, żeby zobaczyć „15 najśliczniejszych kotów na świecie".

Jakiś mężczyzna widzi tę samą kartę w BuzzFeed, ale na ekranie jego komputera nagłówek brzmi: „15 najbardziej uroczych kotów na świecie". Mężczyzna nie wchodzi na stronę.

Inna kobieta wpisuje w wyszukiwarkę: „Czy mój syn jest geniuszem?".

Inny mężczyzna wpisuje w wyszukiwarkę: „Jak mam namówić córkę, żeby się odchudziła?".

Kobieta spędza urlop z sześcioma swoimi najlepszymi przyjaciółkami, które powtarzają w kółko, jak się świetnie bawią. Kobieta wpisuje ukradkiem w wyszukiwarkę: „Samotność z dala od męża".

Mężczyzna, mąż kobiety z poprzedniego akapitu, spędza urlop z sześcioma swoimi najlepszymi przyjaciółmi. Wpisuje ukradkiem w wyszukiwarkę: „Jak poznać, że żona cię zdradza?".

Niektóre z tych danych zawierają informacje, których w innych okolicznościach nie podalibyśmy nikomu. Jeżeli je zsumujemy i zapewnimy internautom anonimowość tak, abyśmy tych lęków, pragnień i postaw nie mogli przyporządkować konkretnym osobom, a potem poddamy dane analizie naukowej, to otrzymamy nowy obraz człowieka – jego zachowań, pragnień i natury. Prawdę powiedziawszy, chociaż nie chciałbym brzmieć pompatycznie, uważam, że nowe informacje, coraz szerzej dostępne w dzisiejszej epoce cyfrowej, radykalnie pogłębią naszą wiedzę o człowieku. Dzięki mikroskopowi okazało się, że w kropli wody ze stawu możemy znaleźć więcej, niż nam się wydaje. Dzięki teleskopowi przekonaliśmy się, że na nocnym niebie widać znacznie więcej, niż sądzimy. A nowe, cyfrowe dane pokazują dzisiaj, że współczesne społeczeństwo nie jest takie, jakim się nam wydaje. Internet może okazać się mikroskopem albo teleskopem

naszych czasów – pozwala bowiem dokonywać ważnych, może nawet przełomowych odkryć.

Z tego rodzaju deklaracjami wiąże się jeszcze jedno ryzyko – że zabrzmią one nie tylko pompatycznie, ale też „trendy". Wiele osób wygłaszało szumne oświadczenia dotyczące potęgi big data, nie miało jednak żadnych dowodów na poparcie swoich tez.

Była to woda na młyn sceptyków – których również nie brakuje – rezygnujących z poszukiwań wielkich zbiorów danych. „Nie twierdzę, że w big data brakuje informacji – napisał eseista i statystyk Nassim Taleb. – Jest ich bardzo dużo. Problem, i to jest zasadnicza kwestia, polega jednak na tym, że igła kryje się w nieustannie rosnącym stogu siana"[11].

Dlatego jednym z głównych celów tej książki jest przedstawienie brakujących dowodów na to, jak można wykorzystać big data – czyli jak można dziś znajdować igły, jeśli zaakceptujemy powyższe porównanie, w coraz większych stogach siana. Mam nadzieję przedstawić wystarczająco dużo przykładów dowodzących, że big data maluje nowy obraz psychiki i zachowania człowieka, dzięki czemu możemy już dziś dojrzeć zarysy naprawdę rewolucyjnego odkrycia.

„Zaraz, zaraz, Seth – mówicie może w tej chwili. – Obiecujesz rewolucję. Używając poetyckiego języka, rozwodzisz się na temat wielkich nowych zbiorów danych, chociaż do tej pory wykorzystałeś wszystkie te rzekomo zdumiewające, niesamowite, zapierające dech w piersi i przełomowe informacje właściwie tylko po to, żeby powiedzieć dwie rzeczy: że w Ameryce mieszka mnóstwo rasistów, a ludzie, zwłaszcza mężczyźni, zawyżają liczbę swoich kontaktów seksualnych".

Przyznaję, że nowe dane czasami po prostu potwierdzają to, co jest skądinąd oczywiste. Jeżeli jednak sądzicie, że przytoczone wyżej ustalenia brzmią banalnie, to zaczekajcie, aż dojdziecie do rozdziału czwartego, gdzie na podstawie wyszukiwań w Google'u przedstawiam jasne i nieodparte dowody na to, że mężczyźni są bardzo niepewni siebie i przewrażliwieni na punkcie – uwaga, napięcie rośnie… – wielkości swojego penisa.

Wydaje mi się bowiem, że czasem warto potwierdzić prawdziwość jakiejś ogólnie przyjętej tezy, na której słuszność nie mieliśmy wcześniej dowodów – gdyż przypuszczenia to jedno, a dowody drugie. Ale jeśli zbiór big data mógłby tylko umacniać nas w przyjętych już przekonaniach, to nie byłoby w nim nic rewolucyjnego. Na szczęście big data potrafi znacznie więcej – i co rusz napawa mnie przekonaniem, że świat wygląda całkowicie inaczej, niż sądziłem. Oto kilka egzemplifikacji, które okażą się może bardziej zaskakujące.

Mogłoby się wydawać, że główną przyczyną rasizmu jest ubóstwo i niepewność jutra, dlatego można naturalnie podejrzewać, że nastroje rasistowskie przybierają na sile, gdy ludzie tracą pracę. W rzeczywistości jednak liczba rasistowskich wyszukiwań czy nowych członków organizacji Stormfront nie zwiększa się wcale wraz ze wzrostem bezrobocia[12].

Można też pomyśleć, że stanów lękowych najczęściej doświadczają wykształceni tak dobrze, że aż przeintelektualizowani mieszkańcy wielkich miast. Stereotyp miejskiego neurotyka jest powszechnie znany. Lecz najwyraźniej przeżywających stany lękowe autorów wyszukiwań w Google'u[13], na przykład o takiej treści jak „symptomy stanów lękowych" czy „pomoc w stanach lękowych", najczęściej obserwujemy więcej tam, gdzie mieszkają ludzie gorzej wykształceni, średnia dochodów jest niższa, a większa część populacji żyje na terenach niezurbanizowanych. Tak więc liczba wyszukiwań dotyczących neurozy okazuje się wyższa w wiejskiej, północnej części stanu Nowy Jork niż w wielkim mieście, od którego ten stan bierze swą nazwę.

Można by sądzić, że po ataku terrorystycznym, w którym giną dziesiątki albo setki ludzi, następuje coś w rodzaju epidemii depresji czy nerwic na tle lękowym. Terroryzm ma w nas przecież wzbudzać grozę niejako z definicji. Przyjrzałem się więc dotyczącym depresji wyszukiwaniom w Google'u. Sprawdziłem, o ile w USA wzrastała ich liczba po kilku dniach, tygodniach i miesiącach od każdego większego ataku terrorystycznego w Europie czy Ameryce, który przeprowadzono po 2004 roku. Nie wzrastała. Wcale.

Można też podejrzewać, że ludzie szukają dowcipów w sieci głównie wtedy, kiedy jest im smutno[14]. Wielu największych mędrców utrzymywało, że w śmiechu poszukujemy ucieczki od bólu. Od dawna sądzono, że śmiech pozwala nam poradzić sobie z rozgoryczeniem, cierpieniem i nieuniknionymi rozczarowaniami, jakie niesie życie. Jak twierdził Charlie Chaplin, śmiech przynosi ulgę, uśmierza, a nawet powstrzymuje ból.

Tymczasem najrzadziej wyszukujemy dowcipy w poniedziałki, kiedy jak sami twierdzimy, jesteśmy w najgorszych nastrojach – a także w pochmurne i deszczowe dni. Liczba wyszukiwanych dowcipów spada także drastycznie po każdej większej tragedii, takiej jak eksplozja dwóch bomb przy trasie bostońskiego maratonu w 2013 roku, w której wyniku zginęły trzy osoby, a kilkaset odniosło rany. Wydaje się więc, że chętniej szukamy dowcipów, gdy układa nam się w życiu dobrze, a nie źle.

Czasami nowy zbiór danych ujawnia jakieś ludzkie zachowania, pragnienia albo obawy, które same nigdy nie przyszłyby mi do głowy. W tej kategorii mieszczą się między innymi nasze upodobania seksualne. Na przykład – czy wiedzieliście, że w Indiach najczęściej wyszukiwana fraza, zaczynająca się od słów „Mój mąż chce…", brzmi: „Mój mąż chce, żebym karmiła go piersią"?[15]. Jest to hasło spotykane znacznie częściej w Indiach niż gdziekolwiek indziej, a treści wizualne przedstawiające kobiety, które karmią piersią dorosłych mężczyzn[16], w Indiach i Bangladeszu wyszukiwane są czterokrotnie częściej niż w jakimkolwiek innym kraju na świecie. Na pewno nigdy bym się tego nie domyślił, gdybym nie poznał danych z Internetu.

Poza tym o ile to, że mężczyźni mają obsesję na punkcie wielkości swojego penisa, nie jest może zbyt zaskakujące, o tyle największe cielesne zmartwienie kobiet, zgodnie z tym, co mówi Google, okazuje się naprawdę nieoczekiwane. Opierając się na nowych danych, możemy powiedzieć, że damskim ekwiwalentem lęków mężczyzn, którzy sądzą, że natura obdarzyła ich zbyt małym przyrodzeniem, jest – i tu napięcie znów wzrasta… – podzielana przez wiele kobiet obawa, że ich wagina wydziela niemiły zapach. Panie wpisują w wyszukiwarkę pytania wyrażające lęki związane ze swoimi genitaliami równie często jak

mężczyźni[17], którzy mają kompleksy na punkcie swoich narządów płciowych, kobiety jednak najczęściej martwią się, że pochwa nieładnie pachnie i chciałyby się dowiedzieć, jak mogą temu zaradzić. Nie miałem o tym pojęcia, dopóki nie przeanalizowałem danych z sieci.

Czasami nowe dane pokazują też różnice kulturowe, których nigdy nawet nie brałem pod uwagę. Przykład: różne sposoby, w jakie reaguje mężczyzna na wiadomość, że jego żona jest w ciąży. W Meksyku do najczęściej wpisywanych w wyszukiwarkę fraz powiązanych z hasłem „moja żona jest w ciąży" należy „frases de amor para mi esposa embarazada"[18] („najserdeczniej pozdrawiam moją ciężarną żonę") albo „poemas para mi esposa embarazada" („strofy dla ciężarnej żony"). W Stanach Zjednoczonych do najczęściej powtarzających się wpisów zaliczamy „moja żona jest w ciąży i co teraz będzie" albo „moja żona zaszła, co mam robić".

Ale *Wszyscy kłamią* jest czymś więcej niż zbiorem ciekawych faktów czy wyników jednorazowych badań, chociaż napotkacie ich tu wiele. Metodologia, którą się posługuję, z czasem będzie jeszcze skuteczniejsza; na razie jest tak nowa, że muszę najpierw przedstawić kilka zasad, na jakich się opiera, i wyjaśnić, na czym polega jej przełomowy charakter. Jednocześnie wskażę ograniczenia big data.

Entuzjazm dla rewolucyjnego potencjału nowych danych bywał dotąd niewłaściwie lokowany. Większość osób urzeczonych big data zachwyca się tym, jak gigantyczne bywają zbiory informacji, przy czym obsesja na punkcie ich wielkości nie jest niczym nowym. Zanim powstały Google, Amazon, Facebook i samo pojęcie big data, w 1977 roku w Dallas, w Teksasie, odbyła się konferencja pod hasłem *Duże i złożone zbiory danych*. Wspomina ją Jerry Friedman, profesor statystyki na Uniwersytecie Stanforda, mój kolega po fachu, którego poznałem w okresie pracy w Google'u. Oto wstaje pewien znakomity statystyk, żeby zacząć swoje wystąpienie, i mówi, że zebrał zdumiewające, niebywałe pięć gigabajtów danych. Po nim podnosi się następny uczony i zaczyna w te słowa: „Mój poprzednik mówił o gigabajtach. To jeszcze nic. Ja mam terabajty". Innymi słowy, uczeni skupiali się na tym, jak dużo informacji można zgromadzić, a nie na tym, co można by z nimi zrobić albo na jakie pytania odpowiedzieć z ich wykorzystaniem.

„Wydało mi się wtedy zabawne – mówi Friedman – że ludzie chcieli sobie zaimponować tym, jak wielkimi zbiorami danych dysponują. Zdarza się to do dziś”[19].

Zbyt wielu współczesnych analityków gromadzi olbrzymie zbiory informacji po to, żeby powiedzieć nam coś bardzo mało ważnego – jak na przykład to, że koszykarska drużyna Knicks cieszy się popularnością w Nowym Jorku. Zbyt wiele firm dosłownie tonie w danych. Mają mnóstwo terabajtów informacji, lecz dokonują niewielu ważnych odkryć. Tymczasem moim zdaniem rozmiary zbioru danych często się przecenia. Istnieje na to subtelne, lecz ważne wytłumaczenie. Im bardziej spektakularny efekt, tym mniejsza liczba obserwacji jest potrzebna, aby go dostrzec. Gorącej kuchenki wystarczy dotknąć raz, żeby się przekonać, że jest niebezpieczna, lecz trzeba wypić może nawet tysiące filiżanek kawy, aby zdać sobie sprawę, że dostajemy po niej bólu głowy. Która z powyższych nauczek jest ważniejsza? Oczywiście ta z kuchenką, bo ze względu na siłę jej rażenia wnioski nasuwają się bardzo szybko, mimo że do ich wyciągnięcia potrzebujemy niewielu obserwacji.

Z tej przyczyny najlepsze firmy analityczne często tną swoje dane. W Google’u na przykład pierwszorzędne decyzje podejmuje się jedynie na bazie skromnej próbki dostępnych materiałów źródłowych[20]. Nie zawsze musimy mieć ich całe tony, żeby dokonać jakiegoś ważnego odkrycia – potrzeba nam po prostu odpowiednich danych. Wyszukiwania w Google’u okazują się dla nas takie cenne nie dlatego, że jest ich bardzo wiele, tylko dlatego, że ludzie są w nich tak szczerzy. Okłamujemy znajomych, kochanków, lekarzy, ankiety i samych siebie, ale w wyszukiwarce Google’a jesteśmy gotowi podzielić się ze światem każdą wstydliwą dla nas informacją, na przykład o deficycie seksu w małżeństwie, kłopotach ze zdrowiem psychicznym, nękających nas obawach czy niechęci do czarnoskórych współobywateli.

Najważniejsze jest to, że aby wycisnąć z big data coś ciekawego, należy zadawać odpowiednie pytania. Tak jak nie da się skierować teleskopu w nocne niebo byle gdzie i wymagać, by odkrył wam Plutona, nie można też po prostu ściągnąć sobie na twardy dysk całej masy danych i oczekiwać, że one same odsłonią nam sekrety ludzkiej natury.

Powinniśmy szukać w obiecujących miejscach – na przykład w Indiach wśród wpisów, które zaczynają się od słów „Mój mąż chce…”.

Wszyscy kłamią pokaże, jak najefektywniej posługiwać się big data, i szczegółowo wyjaśni, dlaczego wielki zbiór danych może okazać się naprawdę potężnym narzędziem badawczym. Po drodze poznacie odpowiedzi, które mnie i innym analitykom udało się już uzyskać, między innymi na takie pytania, jak:

- Jaki jest odsetek homoseksualistów wśród męskiej populacji Stanów Zjednoczonych?
- Czy reklama jest skuteczna?
- Dlaczego American Pharoah był tak dobrym koniem wyścigowym?
- Czy media są stronnicze?
- Czy pomyłki freudowskie zdarzają się naprawdę?
- Kto oszukuje na podatkach?
- Czy uczelnia, na której kończysz studia, ma znaczenie dla twojego sukcesu życiowego?
- Czy można przewidzieć zmiany na giełdzie papierów wartościowych?
- Gdzie najlepiej wychowywać dzieci?
- Co sprawia, że jakaś historia w sieci staje się zaraźliwa w tym sensie, że dzieli się nią ze sobą duża liczba osób?
- O czym powinieneś rozmawiać na pierwszej randce, jeżeli chcesz doprowadzić do drugiej?

I wiele, wiele innych.

Ale zanim do tego wszystkiego dojdziemy, musimy zadać sobie jedno nieco bardziej podstawowe pytanie: po co w ogóle są nam potrzebne bazy danych? Żeby uzyskać odpowiedź, przedstawię wam moją babcię.

CZĘŚĆ PIERWSZA

ZBIORY DANYCH, DUŻE I MAŁE

ROZDZIAŁ PIERWSZY
NASZA OMYLNA INTUICJA

Jeśli masz trzydzieści trzy lata i na kilka kolejnych obiadów z okazji Święta Dziękczynienia przychodziłeś samotnie, przy stole na pewno pojawi się w końcu temat wyboru odpowiedniej partnerki dla ciebie. I oczywiście właściwie każdy będzie miał w tej kwestii coś do powiedzenia.

– Sethowi potrzeba wariatki, bo sam jest wariatem – mówi moja siostra.

– To ty jesteś wariatka! Trzeba mu normalnej dziewczyny, wtedy oboje będą się uzupełniać – twierdzi mój brat.

– Seth nie jest wariatem – odzywa się moja mama.

– Ale z ciebie wariatka! Oczywiście, że Seth jest wariatem – wtrąca ojciec.

Nieoczekiwanie głos zabiera moja nieśmiała, łagodna babcia, która milczała dotąd w czasie obiadu. Donośne, agresywne nowojorskie głosy milkną i wszystkie spojrzenia kierują się na tę drobną starszą panią o krótkich jasnych włosach, mówiącą wciąż jeszcze z lekkim wschodnioeuropejskim akcentem.

– Tobie potrzeba miłej dziewczyny, Seth. Nie za ładnej. Bardzo inteligentnej. I żeby miała dobry kontakt z ludźmi. Żeby była towarzyska, wtedy będziecie mieli ciekawe życie. I żeby miała poczucie humoru, jak ty.

Dlaczego rada staruszki do tego stopnia przykuwa uwagę członków rodziny i budzi ogólny szacunek? No cóż, moja osiemdziesięcioośmioletnia babcia widziała w życiu więcej niż ktokolwiek z nas. Widziała więcej małżeństw niż my, udanych

i nieudanych, i z czasem skatalogowała w głowie wszystkie te elementy, które składają się na dobry związek. Jeżeli chodzi o pytanie dotyczące wyboru właściwej partnerki, to przy naszym stole w Święto Dziękczynienia właśnie moja babcia ma dostęp do największej liczby danych, potrzebnych do udzielenia odpowiedzi. To babcia jest big data.

Chciałbym w tej książce nieco zdemistyfikować analizę danych jako gałąź nauki. Czy komuś się to podoba, czy nie, informacja odgrywa coraz ważniejszą rolę w naszym życiu, a rola ta wkrótce stanie się jeszcze większa. W gazetach codziennych mamy już duże działy poświęcone analizie danych, a współczesne przedsiębiorstwa zatrudniają całe zespoły analityków. Inwestorzy wydają dziesiątki milionów dolarów na start-upy, jeśli tylko są w stanie przechowywać dzięki nim więcej informacji. Nawet jeśli nigdy nie uczyliśmy się analizy regresji albo obliczania przedziału ufności, wszędzie mamy do czynienia z informacją – w tym, co czytamy, na spotkaniach biznesowych albo w plotkach zasłyszanych przy biurowych zbiornikach z wodą, kiedy idziemy się napić.

Ten stan rzeczy niepokoi wielu ludzi. Informacja ich onieśmiela, łatwo się gubią i tracą orientację w świecie liczb. Wydaje im się, że ilościowe rozumienie świata jest dobre wyłącznie dla kilku genialnych mózgowców, nie dla nich, i kiedy tylko napotykają znaki matematyczne, gotowi są przewrócić stronę, zakończyć spotkanie lub zmienić temat rozmowy.

Ja jednak poświęciłem dziesięć lat życia na analizę danych i miałem szczęście pracować z wieloma najwybitniejszymi specjalistami w tej dziedzinie. Nauczyłem się od nich wielu ważnych rzeczy, między innymi tego, że dobra analiza jest mniej skomplikowana, niż to się ludziom wydaje. Powiedziałbym nawet, że najlepsze wyniki można w niej uzyskać, opierając się na intuicji[1].

Jak to możliwe, by analiza danych była intuicyjna? Otóż w gruncie rzeczy chodzi w niej o to, by wykryć pewne wzorce, prawidłowości, i umieć przewidywać, jak jedna zmienna będzie wpływała na drugą. A każdy człowiek robi dokładnie to samo w codziennym życiu.

Zastanówcie się choćby nad tą radą, którą dostałem w kwestii życiowej partnerki. Moja babcia wykorzystała ogromną bazę danych,

zgromadzonych w jej umyśle w ciągu blisko stuletniego życia – czerpała swoją wiedzę z opowieści zasłyszanych od członków rodziny, przyjaciół i znajomych. Analizę ograniczyła jednak do kilku takich przykładowych związków, w których mężczyzna odznaczał się wieloma podobnymi cechami, jakie mam ja – wrażliwością, skłonnością do samotnictwa czy poczuciem humoru. Następnie skupiła się na kluczowych cechach kobiet z takich małżeństw, zastanawiając się, czy i w jakim stopniu czyniły je one dobrymi, miłymi, inteligentnymi i atrakcyjnymi żonami. Następnie zestawiła kluczowe cechy kobiece z najważniejszą własnością związku, pytając, czy był udany, na końcu przedstawiła zaś wyniki tego procesu myślowego. Innymi słowy, dostrzegła określone prawidłowości i przewidziała, w jaki sposób jedna zmienna będzie wpływać na drugą. Moja babcia jest analitykiem danych.

I wy też nim jesteście. Jak byliście mali, spostrzegliście, że kiedy płaczecie, mama zaczyna się wami interesować. To analiza danych. Kiedy dorośliście, zorientowaliście się, że jeśli za często narzekacie, znajomi nie bardzo mają chęć na wasze towarzystwo. To również jest przykład myślenia analitycznego. Zauważyliście ponadto, że kiedy ludzie rzadziej z wami przestają, popadacie w gorszy nastrój, a w gorszym nastroju jesteście mniej przyjaźnie nastawieni do świata. I wtedy znajomi jeszcze mniej mają ochotę się z wami zadawać. Analiza danych. Analiza danych. Analiza danych.

Przekonałem się, że ponieważ jest ona tak naturalna, najciekawsze badania spod znaku big data może zrozumieć właściwie każdy inteligentny człowiek. Jeżeli nie, to coś jest zapewne nie w porządku z tymi badaniami, a nie z nim.

Chcecie dowodu, że najlepsza analiza danych opiera się na ogół na intuicji? Natknąłem się niedawno na wyniki być może najważniejszych studiów analitycznych, jakie przeprowadzono w ostatnich kilku latach, a przy tym zapewne najbardziej intuicyjnych, jakie kiedykolwiek widziałem. Zastanówcie się nie tylko nad ich doniosłym znaczeniem, ale też nad tym, w jak naturalny sposób, przypominający metodę mojej babci, wyciągnięto z nich wnioski.

Badania, o których mówię, przeprowadził zespół naukowców z Columbia University i z firmy Microsoft. Uczeni chcieli ustalić, jakie

symptomy mogą pomóc przewidzieć ryzyko wystąpienia raka trzustki[2]. Szanse na wyleczenie tej choroby są bardzo niskie, wynoszą około trzech procent, ale jej wczesne wykrycie może je podwoić.

Jaką metodę zastosowali badacze? Wykorzystali dane z wpisów dziesiątków tysięcy użytkowników Binga, wyszukiwarki Microsoftu. To, że u danego internauty zdiagnozowano niedawno tę chorobę, stwierdzano na podstawie jednoznacznych wpisów z jego strony, takich jak „właśnie zdiagnozowano u mnie raka trzustki" albo „powiedziano mi, że mam raka trzustki, czego mogę w związku z tym się spodziewać?".

Później uczeni przyjrzeli się wyszukiwaniom dotyczącym ogólnych problemów zdrowotnych i porównali wpisy niewielkiej liczby tych internautów, u których po pewnym czasie stwierdzono raka, z wpisami tych, którzy na niego nie zapadli. Właśnie w ten sposób poszukiwali symptomów sygnalizujących, że za kilka tygodni czy miesięcy dana osoba zachoruje.

Rezultaty były zdumiewające. Okazało się, że bóle pleców połączone z żółknięciem skóry to objaw nowotworu trzustki, choć same bóle pleców raczej wykluczały tę chorobę. Zapowiadała ją też niestrawność, połączona z bólami żołądka, mimo że sama niestrawność nie świadczyła o żadnym zagrożeniu. Zespołowi badawczemu udało się wytypować od pięciu do piętnastu procent przypadków zachorowań niemal bezbłędnie. Być może nie wygląda to na zbyt imponujący wynik, ale jeżeli chorujesz na raka wspomnianego narządu, to nawet dziesięć procent szans na prawdopodobne zwiększenie możliwości wyleczenia jest czystym darem niebios.

Rzecz jasna, artykuł, w którym szczegółowo opisano powyższe badania, trudno byłoby w pełni zrozumieć niespecjaliście. Zawiera sporo technicznego żargonu – mówi na przykład o teście Kołmogorowa-Smirnowa, a muszę tutaj przyznać, że już sam zapomniałem, na czym on dokładnie polega (pamiętam tylko, że jest stosowany do porównywania rozkładów jednowymiarowych cech statystycznych i pozwala określić, czy przyjęty model da się dopasować do konkretnego zbioru danych).

Zwróćcie jednak uwagę, jak naturalne i intuicyjne są te zdumiewające badania na poziomie najbardziej podstawowym. Uczeni przyjrzeli się szerokiemu spektrum przypadków medycznych i próbowali połączyć symptomy z konkretną chorobą. Wiecie, kto jeszcze używa takiej metody, starając się ustalić, czy człowiek jest na coś chory? Mężowie i żony, matki i ojcowie, pielęgniarki i lekarze. Opierając się na doświadczeniu i wiedzy, usiłujemy połączyć gorączkę, ból głowy, katar i ból brzucha z rozmaitymi chorobami. Mówiąc wprost, naukowcy z Uniwersytetu Columbia i firmy Microsoft przeprowadzili przełomowe badania, stosując naturalną, oczywistą metodę, której używamy wszyscy, kiedy chcemy postawić jakąś diagnozę.

Ale zaraz, zaraz. Zwolnijmy tempo. Jeżeli najlepsze analizy danych przeprowadza się często w sposób naturalny i intuicyjny, jak w tej chwili twierdzę, to powstaje podstawowe pytanie o wartość big data. Jeśli ludzie są z natury analitykami danych, jeżeli analiza jest intuicyjna, to po co nam komputery i specjalistyczne oprogramowanie? Po co testy Kołmogorowa-Smirnowa? Czy nie możemy po prostu działać na podstawie intuicji, jak moja babcia albo lekarze i pielęgniarki?

W tym miejscu dochodzimy do sporu, który przybrał na sile zwłaszcza po wydaniu bestsellera Malcolma Gladwella *Błysk! Potęga przeczucia*[2*], wynoszącego pod niebiosa czarodziejską moc naszych intuicji i instynktów. Autor opowiada w nim o ludziach, którzy opierając się wyłącznie na intuicji, potrafią rozpoznać sfałszowaną rzeźbę, przewidzieć, czy tenisista popełni błąd, zanim jeszcze w ogóle dotknie piłki, i ile dany klient gotów jest wydać pieniędzy. Bohaterowie tej książki nie przeprowadzają analizy regresji, nie obliczają przedziałów ufności i nie stosują testów Kołmogorowa-Smirnowa, ale w większości wypadków zdumiewająco odgadują prawdę. Wiele osób intuicyjnie wspiera taką apologię intuicji, jaką przeprowadza Gladwell, ponieważ sami ufają swoim przeczuciom i wrażeniom. Wielbiciele *Błysku* będą zapewne unosić się nad mądrością mojej babci, która udziela mi rad na temat związków damsko-męskich bez pomocy komputerów, ale już z mniejszym entuzjazmem podejdą do wykorzystujących te urządzenia moich czy innych badań, opisanych w tej książce. Jeżeli zatem to komputerowa, a nie babcina analiza big

data stanowi autentyczną rewolucję, musi ona udowodnić, że jest potężniejsza niż niewspomagana maszynowo ludzka intuicja, która, jak słusznie wskazał Gladwell, rzeczywiście bywa zdumiewająca.

Badania uczonych z Columbii i Microsoftu stanowią dobry przykład tego, jak rygorystyczna analiza danych i komputery mogą dać nam coś, czego nigdy byśmy nie zdobyli, polegając wyłącznie na intuicji. Zresztą w tym wypadku liczy się akurat także wielkość zbioru. A człowiekowi brakuje czasem odpowiedniego doświadczenia, by intuicja niewspomagana z zewnątrz mogła mu wystarczyć. Jest mało prawdopodobne, że ty, twoi bliscy znajomi albo krewni znacie na tyle dużo przypadków raka trzustki, by dostrzec jego zależność od niestrawności powiązanej z bólami brzucha i niezależność od samej niestrawności. Jest wręcz nieuniknione, że wraz z powiększaniem się zbioru danych z Binga uczeni zauważą w związku z tą i innymi chorobami znacznie więcej różnych subtelnych relacji symptomatycznych, które mogłyby ujść nawet uwagi lekarzy.

Co więcej, chociaż intuicja zwykle podsuwa nam stosunkowo prawdziwy obraz świata, często bywa mało precyzyjna. Potrzebujemy danych, żeby wyostrzyć obraz. Weźcie na przykład pod uwagę oddziaływanie pogody na nastroje. Sądzicie zapewne, że ludzie częściej czują się przygnębieni, kiedy na zewnątrz jest minus dziesięć stopni, nie plus dwadzieścia pięć. Bo rzeczywiście tak jest. Ale chyba nie zdajecie sobie sprawy, jak kolosalne znaczenie może mieć różnica temperatur. Szukałem kiedyś korelacji między wyszukiwaniami w Google'u treści dotyczących depresji a miejscem zamieszkania, uwzględniając szeroki wachlarz czynników, takich jak warunki ekonomiczne, poziom wykształcenia i religijność internautów. Zimny klimat pobił je wszystkie na głowę[3]. W okresie zimowym w ciepłych miejscach, takich jak Honolulu na Hawajach, odnotowujemy około czterdziestu procent mniej wpisów na temat depresji niż w chłodnych rejonach kraju, jak powiedzmy, Chicago w stanie Illinois. Na ile taka obserwacja jest istotna? Otóż całkiem optymistyczna ocena działania antydepresantów mówi, że najskuteczniejsze z nich zmniejszają ryzyko depresji najwyżej o dwadzieścia procent. Sądząc zatem po wynikach z Google'a, przeprowadzka z Chicago do Honolulu byłaby co najmniej

dwa razy skuteczniejsza niż nawet najlepsze lekarstwo na zimowy blues[3*].

Czasami intuicja, jeśli nie stoi za nią staranna analiza komputerowa, myli się na całej linii, kiedy zaślepiają nas osobiste doświadczenia i przesądy. Wprawdzie babcia potrafi wykorzystać dziesięciolecia doświadczeń i udzielić mi dobrej rady co do odpowiedniej dla mnie partnerki, ale już jej poglądy na to, co zapewnia trwałość związkowi dwojga ludzi, wydają się wątpliwe – twierdziła na przykład bowiem wielokrotnie, że trzeba mieć wspólnych przyjaciół. Zdaniem babci był to kluczowy czynnik w sukcesie jej małżeństwa, bo większość ciepłych wieczorów spędzała ze swoim mężem, czyli moim dziadkiem, w małym ogródku na nowojorskim Queensie, siedząc na krzesełkach turystycznych i plotkując z grupką najbliższych sąsiadów.

Zabrzmi to pewnie tak, jak gdybym chciał teraz kopać dołki pod własną babcią, ale analiza danych pokazuje, że jej teoria nie jest prawdziwa. Zespół informatyków przeanalizował niedawno największą bazę danych, jaką kiedykolwiek zebrano na temat związków międzyludzkich – Facebooka[4]. Naukowcy przyjrzeli się dużej liczbie par, które w pewnym momencie swojego życia pozostawały ze sobą „w związku". Niektóre go utrzymały, inne się rozeszły. Ustalono jednak, że jeśli para ma wielu wspólnych znajomych, to najprawdopodobniej w końcu się rozstanie. Być może wspólne spędzanie czasu co wieczór z partnerem czy partnerką i ciągle tą samą grupą osób wcale nie jest dobre, a odrębne kręgi towarzyskie pomagają wzmocnić bliskie więzi.

Jak widzimy, czysta intuicja, jeśli trzymać się z dala od komputera i iść za jej głosem, może niekiedy przynieść zdumiewające rezultaty, może też jednak błądzić. Moja babcia wpadła zapewne w pułapkę kognitywną, bo wszyscy mamy skłonności do wyolbrzymiania znaczenia osobistych doświadczeń. W żargonie analityków „ważymy" nasze dane, czyli przywiązujemy zdecydowanie zbyt dużą wagę do jednego z elementów zbioru – nas samych.

Babcia tak mocno skupiała się na wieczornych pogaduszkach z dziadkiem i przyjaciółmi, że nie myślała wystarczająco dużo o innych małżeństwach. Zapomniała wziąć pod uwagę swojego szwagra i jego żonę, którzy spotykali się prawie co wieczór na plotki w niewielkim,

stale takim samym kręgu znajomych, lecz często się kłócili i w końcu rozwiedli. Zapomniała uwzględnić też moich rodziców, czyli swoją córkę i zięcia, którzy często spędzali czas osobno – tata chodził do klubu jazzowego albo na mecze z kolegami, a mama wybierała się do restauracji albo kina ze swoimi znajomymi: mimo to pozostają szczęśliwym małżeństwem do dziś.

Jeśli polegamy na intuicji, na manowce może nas zaprowadzić także głęboko ludzka fascynacja tym, co dramatyczne. Mamy skłonności do przeceniania znaczenia wszystkiego, co nam zapada w pamięć. Na przykład odpowiadając na pytania w ankietach, ludzie zawsze częściej wymieniają tornada jako powszechniejszą przyczynę śmierci niż astmę[5], chociaż w USA ta choroba przyczynia się do zgonów ponad siedemdziesiąt razy częściej niż huraganowy wiatr[6]. Po prostu śmierć z powodu astmy niczym się nie wyróżnia, nie trafia więc na pierwsze strony gazet, a śmierć spowodowana przez tornado – i owszem.

A zatem jeśli polegamy wyłącznie na tym, co słyszymy albo czego doświadczamy osobiście, to często się mylimy co do tego, jak naprawdę funkcjonuje świat. I chociaż rzetelna analiza danych nierzadko opiera się na intuicji, to jej wyniki równie często bywają kontrintuicyjne. Można by powiedzieć, że analiza jak gdyby nasyca sterydami naturalny, intuicyjny proces myślowy, polegający na znajdowaniu pewnych prawidłowości oraz ich sensów, i pokazuje nam, że świat jest urządzony zupełnie inaczej, niż sądziliśmy. Przekonałem się o tym na przykład, kiedy zacząłem szukać odpowiedzi na pytanie, jakie czynniki sprawiają, że jedni koszykarze osiągają sukcesy w swojej dyscyplinie sportu, a inni nie.

Gdy byłem dzieckiem, miałem jedno i tylko jedno marzenie. Chciałem zostać ekonomistą i analitykiem danych. No nie. Żartuję. Strasznie chciałem zostać zawodowym koszykarzem i pójść w ślady mojego idola, Patricka Ewinga[7], prawdziwego gwiazdora, centra New York Knicks.

Czasami wydaje mi się, że w każdym analityku siedzi dzieciak, który stara się odgadnąć, dlaczego nie spełniły się jego marzenia z młodości –

a zatem nic dziwnego, że niedawno spróbowałem zbadać, jakie należy mieć cechy, by trafić do NBA. Wyniki okazały się zaskakujące. Prawdę mówiąc, pokazują one po raz kolejny, jak poprawna analiza danych może zmienić nasze poglądy na świat i jak przeciwintuicyjne mogą się okazać jej rezultaty.

Podstawowe pytanie, które sobie zadałem, brzmiało: czy masz większe szanse na karierę w NBA, jeśli wychowywałeś się w ubogiej rodzinie, czy raczej jeśli pochodzisz z klasy średniej?

Większość z nas wskazałaby chyba ubóstwo. Ludzie na ogół sądzą, że dorastanie w trudnych warunkach, powiedzmy w blokach komunalnych, bez ojca, tylko pod opieką samotnej nastoletniej matki, pomaga wyzwolić w człowieku energię, potrzebną, żeby osiągnąć szczyty w tym niełatwym sporcie, wymagającym wyjątkowego ducha rywalizacji.

Taką opinię wyraził na przykład licealny trener koszykówki z Filadelfii William Ellerbee w wywiadzie, jakiego udzielił pismu „Sports Illustrated". „Dzieciaki z małych miasteczek i bogatych przedmieść grają na ogół dla zabawy – stwierdził. – A młodzież z dużych miast i dzielnic biedy traktuje koszykówkę jak sprawę życia i śmierci"[8]. Niestety, ja wychowałem się w tak zwanej pełnej rodzinie w małym miasteczku w New Jersey. Natomiast LeBron James, najlepszy gracz mojego pokolenia, przyszedł na świat w ubogim domu w Akron w stanie Ohio, kiedy jego matka miała szesnaście lat.

Co więcej, ankieta internetowa, jaką przeprowadziłem na ten temat[9], zdawała się wskazywać, że większość Amerykanów myśli tak samo jak Ellerbee i ja – że prawie wszyscy gracze w NBA wychowywali się w biedzie.

Tylko czy ta powszechna opinia jest prawdą?

Spójrzmy na dane. Nie istnieje żadna ich kompletna baza, przedstawiająca warunki społeczno-ekonomiczne, w których przyszło dorastać graczom NBA. Lecz jako detektywi informacyjni, używający danych z wielu różnych źródeł – na przykład z basketball-reference.com, ancestry.com, powszechnego spisu ludności Stanów Zjednoczonych i tak dalej – możemy się dowiedzieć, jakie środowisko

rodzinne najbardziej sprzyja karierze koszykarskiej. Proszę zwrócić uwagę, że analiza ta korzysta z bardzo różnych źródeł danych, obszerniejszych i skromniejszych, online i offline. Nowe skarbnice cyfrowe oczywiście są ekscytujące, lecz dobry analityk nie zlekceważy źródeł staroświeckich, jeśli tylko mogą być pomocne. Najlepszy sposób uzyskania odpowiedzi na takie czy inne pytanie to odwołać się do wszystkich dostępnych danych.

Pierwsza ważna informacja dotyczy miejsca urodzenia gracza. Odnotowałem więc najpierw, ilu białych i czarnych mężczyzn przyszło na świat w latach osiemdziesiątych we wszystkich hrabstwach USA z osobna. Potem sprawdziłem, ilu z nich trafiło do NBA, i zestawiłem te liczby z przeciętnym wynagrodzeniem w danym hrabstwie. Sprawdzałem też demograficzne dane rasowe, ponieważ – i to jest temat na osobną książkę – czarnoskórzy mężczyźni mają czterdziestokrotnie większe szanse na karierę w NBA niż biali.

Dane mówią nam, że jeśli przyszedłeś na świat w hrabstwie zamożnym, to istnieje znacznie większe prawdopodobieństwo, że trafisz do zawodowej ligi koszykówki. I tak na przykład czarnoskóry chłopak urodzony w którymś z najbogatszych hrabstw w Stanach Zjednoczonych ma dwa razy większe szanse na sukces niż czarny z jednego z najuboższych hrabstw. A jeśli chodzi o białych, to szanse chłopaka z bogatego hrabstwa wynoszą o sześćdziesiąt procent więcej.

Sugerowałoby to, że wbrew powszechnej opinii sportowców z ubogich środowisk jest w NBA nieproporcjonalnie mało. Dane te nie są jednak zbyt dokładne, bo wiele zamożnych hrabstw, na przykład Nowy Jork (Manhattan), obejmuje też dzielnice ubogie, jak Harlem. A zatem wciąż jest możliwe, że trudne dzieciństwo sprzyja udanej karierze koszykarskiej. Potrzeba więcej wskazówek, więcej danych.

Dlatego zbadałem również sytuację rodzinną graczy z NBA. Informacje na ten temat czerpałem z prasy i mediów społecznościowych. Była to metoda nad wyraz czasochłonna, ograniczyłem więc swoją analizę do stu urodzonych w latach osiemdziesiątych zawodników afroamerykańskich, którzy zdobywali najwięcej punktów w lidze. W porównaniu ze średnią krajową okazało się o trzydzieści procent mniej prawdopodobne, że czarna

supergwiazda koszykówki będzie dzieckiem samotnej lub nastoletniej matki. A w takim razie dobra sytuacja rodzinna najlepszych graczy NBA również wskazuje na to, że chodzi o ważny czynnik, pomagający im osiągnąć sukces.

Pamiętajmy jednak przy tym, że dane „urodzinowe" z poszczególnych hrabstw czy sytuacja rodzinna ograniczonego zbioru koszykarzy nie zapewniają przecież kompletnych informacji na temat dzieciństwa wszystkich zawodowców w NBA. Z tego powodu nadal nie byłem całkiem przekonany, że rodzice z klasy średniej dostarczają do NBA więcej gwiazd koszykarskich niż rodziny ubogie i niepełne. Im więcej danych zbierzemy, żeby wyklarować tę kwestię, tym lepiej.

Przypomniałem sobie, że jest jeszcze jeden czynnik, dzięki któremu mógłbym pozyskać ważne informacje o interesujących mnie sportowcach. Otóż pewien artykuł autorstwa dwóch ekonomistów, Rolanda Fryera i Stevena Levitta, sugerował, że w wypadku czarnoskórych obywateli USA na status społeczno-ekonomiczny wskazują ich imiona[10]. Fryer i Levitt przestudiowali świadectwa urodzenia w Kalifornii z lat osiemdziesiątych i zauważyli, że wśród Afroamerykanów ubogie, niewykształcone i samotne matki nadają swoim dzieciom wyraźnie inne imiona niż wykształceni rodzice z klasy średniej.

Dzieci z zamożniejszych domów dostają raczej miana pospolite, jak Kevin, Chris czy John, a dzieci z rodzin uboższych i bloków komunalnych stosunkowo rzadkie, jak Knowshon, Uneek czy Breionshay. Afroamerykańska młodzież z ubogich środowisk niemal dwukrotnie częściej nosi imiona oryginalne, wyjątkowe, których nie otrzymują inne dzieci z ich rocznika.

Jak w takim razie wygląda wykaz imion czarnoskórych graczy NBA? Wskazuje na klasę średnią czy raczej na pochodzenie z ubogich rodzin? Okazało się, że koszykarze urodzeni w Kalifornii otrzymują niecodzienne imiona o pięćdziesiąt procent rzadziej niż przeciętny czarny Amerykanin, co statystycznie jest różnicą znaczącą.

Znacie kogoś, kto sądzi, że NBA to liga dla chłopaków z dzielnic biedy? Jeśli tak, powiedzcie mu, żeby uważnie posłuchał najbliższej radiowej transmisji z meczu koszykarskiego. Poproście, żeby policzył,

ile razy Russell mija Dwighta i unikając wyciągniętych ramion Josha, próbuje podać piłkę do czekającego na nią z otwartymi rękami Kevina. Gdyby NBA rzeczywiście była ligą zdominowaną przez graczy z ubogich czarnych rodzin, to ta relacja brzmiałaby całkiem inaczej. Bo słyszelibyśmy w niej o wiele więcej imion takich jak LeBron.

Mamy więc dane z trzech źródeł: wiemy, w jakim hrabstwie urodził się dany gracz, wiemy, czy matki najlepszych strzelców są samotne i ubogie, czy raczej zamężne i z klasy średniej, oraz znamy imiona interesujących nas koszykarzy. Żadne źródło nie jest doskonałe i nie zawiera wszystkiego. Tutaj jednak wszystkie trzy wskazują na jedno i to samo. Wyższy status społeczno-ekonomiczny oznacza większe szanse na grę w NBA. Inaczej mówiąc, powszechna opinia na ten temat jest mylna.

Około sześćdziesięciu procent Afroamerykanów urodzonych w latach osiemdziesiątych miało rodziców, którzy nie byli małżeństwem[11], lecz z moich obliczeń wynika, że rodzice ogromnej większości czarnych urodzonych w tej dekadzie i grających w NBA pozostawali jednak związani węzłem małżeńskim. Oznacza to, że nasza liga zawodowa nie składa się głównie z graczy mających za sobą taką przeszłość jak LeBron James. Większość z nich przypomina wychowanego w pełnej rodzinie Chrisa Bosha z Teksasu, który interesował się i nadal interesuje gadżetami elektronicznymi, albo Chrisa Paula[12], drugiego syna rodziców z klasy średniej z Lewisville w Karolinie Północnej; jego ojciec i matka wystąpili nawet razem z nim i innymi członkami rodziny w jednym z odcinków teleturnieju *Family Feud* [*Familiada*] w 2011 roku.

Celem analityka danych jest zrozumieć świat. Kiedy trafimy na jakiś kontrintuicyjny wynik, możemy kontynuować analizę danych, która pomogłaby nam wyjaśnić, dlaczego świat nie jest taki, jak nam się wydaje. Dlaczego na przykład mężczyźni z klasy średniej mają więcej szans na koszykarski sukces niż ci, którzy pochodzą z ubogich rodzin? Istnieją po temu co najmniej dwa wytłumaczenia.

Pierwsze mówi, że mężczyźni z uboższych domów są niżsi. Uczeni wiedzą od dawna, że dobra opieka zdrowotna w dzieciństwie i dobre odżywianie silnie wpływają na zdrowie w dorosłym życiu. Dlatego przeciętny mężczyzna z krajów dobrze rozwiniętych liczy sobie dzisiaj

cztery cale wzrostu więcej niż sto pięćdziesiąt lat temu[13]. A dane wskazują na to, że Amerykanie z ubogich rodzin[14] są niżsi właśnie dlatego, że mają gorszą opiekę zdrowotną w dzieciństwie i wczesnej młodości i gorzej się odżywiają.

Dane mówią także, jaki wpływ na przebicie się do NBA ma wzrost. Na pewno intuicja podpowiada wam, że wysoki wzrost to atut dla początkującego koszykarza; wystarczy porównać przeciętnego zawodowca na parkiecie z kibicem na widowni – gracze NBA mierzą średnio sześć stóp i siedem cali, a statystyczny Amerykanin ma pięć stóp i dziewięć cali wzrostu[15].

A zatem jak istotne dla koszykarskiego sukcesu są warunki fizyczne? Gracze NBA czasami trochę fantazjują na swój temat, a nie ma kompletnych danych dotyczących rozkładu krzywej wzrostu amerykańskich mężczyzn. Posługując się jednak przybliżonym matematycznym modelem takiej krzywej oraz danymi z NBA, łatwo potwierdzić, że znaczenie warunków fizycznych jest olbrzymie – może nawet większe, niż przypuszczaliśmy. Z moich szacunków wynika, że każdy dodatkowy cal podwaja, z grubsza biorąc, szanse gracza na wejście do zawodowej ligi. I jest to twierdzenie prawdziwe dla wszystkich elementów zbioru. Mężczyzna o wzroście pięciu stóp i jedenastu cali ma dwa razy większe szanse na grę w NBA od mężczyzny o wzroście pięciu stóp i dziesięciu cali, a mężczyzna mierzący sześć stóp i jedenaście cali dwa razy większe niż liczący sześć stóp i dziesięć cali. Wydaje się, że spośród młodych ludzi poniżej sześciu stóp wzrostu do NBA trafia zaledwie jeden na dwa miliony, a spośród tych mierzących ponad siedem w zawodowej lidze koszykówki gra mniej więcej jeden na pięciu[16] – przynajmniej tak wynika z badań prowadzonych przeze mnie i innych analityków.

Proszę zwrócić uwagę, że dane te wyjaśniają, dlaczego moje marzenie o zostaniu gwiazdą koszykówki nigdy się nie spełniło. Wcale nie z tego powodu, że wychowałem się w małym miasteczku, tylko dlatego, że mierzę pięć stóp i dziewięć cali wzrostu i jestem biały, a także zbyt powolny i leniwy. I mało wytrzymały. Na dodatek rzucam

bardzo niecelnie, a kiedy piłka trafia w moje ręce, zdarza mi się dostać ataku paniki.

Drugi powód, dla którego chłopcy z trudnych dzielnic mogą mieć problem z dostaniem się do NBA, jest taki, że czasami brakuje im pewnych umiejętności społecznych. Korzystając z danych dotyczących tysięcy uczniów szkolnych, ekonomiści stwierdzili, że rodzice z klasy średniej[17] na ogół lepiej radzą sobie z wychowaniem dzieci, które dzięki ich opiece są ufne, zdyscyplinowane, cierpliwe, skoncentrowane na osiągnięciu wyznaczonego celu i zorganizowane.

A w jaki sposób nieprzystosowanie społeczne przeszkadza w poza tym dobrze zapowiadającej się indywidualnej karierze koszykarskiej?

Przyjrzyjmy się historii Douga Wrenna, jednego z najlepiej rokujących graczy lat dziewięćdziesiątych. Jego trener uniwersytecki, Jim Calhoun z Uniwersytetu Connecticut, który pracował z wieloma przyszłymi gwiazdami NBA, twierdził, że był to najskoczniejszy koszykarz, jakiego w życiu trenował. Wrenn wychowywał się jednak w trudnych warunkach[18]. Opiekowała się nim samotna matka, a mieszkał na Blood Alley, w jednej z najbardziej niespokojnych dzielnic Seattle. W Connecticut nieustannie ścierał się z osobami ze swojego najbliższego otoczenia. Prowokował innych graczy, kwestionował decyzje trenerów i wbrew przepisom obowiązującym w drużynie nosił na parkiecie zbyt luźne spodenki i koszulki. Miał też kłopoty z prawem – wdawał się w pyskówki z policjantami, raz ukradł buty w sklepie. Cierpliwość Calhouna wreszcie się skończyła, wyrzucił go więc z drużyny.

Wrenn dostał drugą szansę na University of Washington, ale i tam przeszkadzała mu niezdolność do zgodnego współżycia z otoczeniem. Kłócił się z trenerem, ponieważ uważał, że za rzadko wpuszcza go na boisko, i też został w końcu skreślony ze składu. Nie otrzymał żadnej propozycji z NBA, kręcił się po niższych ligach, przeprowadził się z powrotem do matki, a w końcu trafił do więzienia za rozbój. „Moja kariera sportowa jest skończona – powiedział gazecie «Seattle Times» w 2009 roku. – To koniec moich marzeń i ambicji. Doug Wrenn umarł. Znaczy jako koszykarz; ten gościu już nie żyje[19]. To jego koniec”.

A Wrenn miał dość talentu, by zostać nie tylko graczem NBA, ale wielkim, może nawet legendarnym zawodnikiem. Nie udało mu się jednak zapanować nad swoim temperamentem do tego stopnia, żeby utrzymać się choćby w drużynie uniwersyteckiej. Może gdyby miał normalne dzieciństwo, stałby się następnym Michaelem Jordanem.

Jordan oczywiście też był niezwykle skoczny. Odznaczał się przy tym niesłychaną ambicją i imponującym instynktem współzawodnictwa, ale charakterologicznie pod pewnymi względami przypominał Wrenna. Bo Michael Jordan też był trudnym dzieckiem[20]. Relegowano go ze szkoły w wieku dwunastu lat, gdyż często wdawał się w bójki. Od Wrenna różnił się natomiast na pewno pod jednym względem: miał spokojne dzieciństwo i pochodził z klasy średniej. Jego ojciec był kontrolerem jakości w General Electric, matka pracowała w sektorze bankowym[21]. Oboje pomogli mu pokierować karierą.

Życiorys Jordana obfituje w opowieści o tym, jak rodzice wskazywali synowi sposoby uniknięcia różnorodnych pułapek, czyhających na każdy wielki i ambitny talent sportowy[22]. Kiedy małego Michaela wyrzucono ze szkoły, matka zaczęła zabierać go ze sobą do pracy. Nie wolno mu było wychodzić z samochodu, musiał siedzieć na parkingu i czytać książki. Gdy trafił do drużyny Chicago Bulls, rodzice i rodzeństwo odwiedzali go tam na zmianę, żeby pilnować chłopaka i nie dopuścić do niego pokus przychodzących ze sławą i pieniędzmi.

Jego kariera nie skończyła się tak jak kariera Wrenna, czyli krótką wzmianką w „Seattle Times", tylko przemówieniem, które wygłosił po przyjęciu go do Basketball Hall of Fame na oczach milionów telewidzów[23]. Powiedział, że starał się „skupić na tym, co w życiu dobre – jak ludzie cię odbierają, jak ich szanować... Na tym, jak jesteś odbierany publicznie. Jak się zatrzymywać i przez chwilę zastanowić nad tym, co się robi. Nauczyli mnie tego moi rodzice"[24].

Zbiory danych potwierdzają, że Jordan ma całkowitą rację, dziękując ojcu i matce (z klasy średniej), którym udało się stworzyć dobre małżeństwo. Analiza przekonuje bowiem, że w uboższych rodzinach i środowiskach też rodzą się talenty na miarę NBA, które ostatecznie tam jednak nie trafiają. Ci młodzi ludzie mają odpowiednie geny

i ambicję, lecz nie udaje im się tak ukształtować swojej osobowości, by mogli zostać prawdziwymi gwiazdami koszykówki.

I bez względu na to, co nam podpowiada intuicja, życie w tak rozpaczliwych okolicznościach, że koszykówka wydaje się młodym chłopakom „kwestią życia i śmierci", wcale nie pomaga w karierze sportowej. Przypadki takie jak Wrenna tylko to potwierdzają. Dane analityczne też.

W czerwcu 2013 roku LeBron James udzielił wywiadu telewizyjnego po zdobyciu swojego drugiego mistrzostwa NBA (potem zdobył jeszcze trzecie)[25]. „Jestem LeBron James – powiedział. – Pochodzę z Akron w Ohio. Z ubogiej dzielnicy miasta. I teoretycznie rzecz biorąc, właściwie w ogóle nie powinno mnie tu być". Twitter oraz inne media społecznościowe eksplodowały krytyką. Jak ten niebywale utalentowany człowiek, który od śmiesznie wczesnego dzieciństwa uważany był za przyszłość koszykówki, może twierdzić, że miał mniejsze szanse od innych? A przecież prawda wygląda tak, że każdy, kto wywodzi się z trudnego środowiska, ma w życiu do pokonania znacznie więcej przeszkód. Z tego punktu widzenia osiągnięcia Jamesa są jeszcze bardziej wyjątkowe, niż mogą się wydawać na pierwszy rzut oka. Dane analityczne również to potwierdzają.

CZĘŚĆ DRUGA

MOCE BIG DATA

ROZDZIAŁ DRUGI
CZY FREUD MIAŁ RACJĘ?

Zauważyłem niedawno, że ktoś napisał o idącym ulicą mężczyźnie „penistrian". Rozumiecie, prawda? „Penistrian" zamiast „pedestrian" [„przechodzień"]. Natknąłem się na to słowo w wielkim zbiorze literówek, jakie popełniają internauci. Widzimy nieznajomego na ulicy i piszemy „penis". To chyba musi coś znaczyć?

Słyszałem też ostatnio o mężczyźnie, któremu śniło się, że je banana w drodze do ołtarza na zaślubiny ze swoją narzeczoną. Znalazłem tę historię w dużym zbiorze snów, które ludzie zapisują w sieci. Mężczyzna wyobraża sobie, że bierze ślub z kobietą, jedząc falliczny owoc[1]. To pewnie też coś znaczy, prawda?

Czy Sigmund Freud miał rację? Najbardziej uczciwą odpowiedzią na to pytanie, zadawane, odkąd jego teorie zaczęły się cieszyć szerszym zainteresowaniem, byłoby wzruszenie ramion. Dlaczego? Najjaśniej wytłumaczył to Karl Popper, austriacki filozof, który większość życia spędził w Anglii. Popper zasłynął stwierdzeniem, że teorii Freuda nie da się sfalsyfikować. Nie można wykazać ich prawdziwości, tak jak nie można dowieść, że są fałszywe.

Freud mógłby dowodzić, że osoba, która napisała „penistrian", ujawnia być może swoje tłumione pragnienie seksualne. Osoba ta mogłaby jednak odpowiedzieć, że absolutnie niczego nie ujawnia i że po prostu zrobiła niewinną literówkę tego samego rodzaju, jak na przykład „pedaltrian". Mielibyśmy zatem sytuację typu „słowo przeciw słowu". Freud mógłby też powiedzieć, że mężczyzna, któremu śniło się, że je banana w dniu swojego ślubu, myślał podświadomie o penisie, ujawniając w ten sposób, że w rzeczywistości pragnie poślubić mężczyznę. Śniący mógłby jednak utrzymywać, że po prostu

przypadkiem śnił mu się banan i równie dobrze mogłoby mu się przyśnić, że po drodze do ołtarza je jabłko. Jeden mówi jedno, drugi drugie. Nie było żadnego sposobu, by poddać teorię Freuda sprawdzianowi.

Aż do dzisiaj.

Analiza danych falsyfikuje wiele twierdzeń Freuda i jest w stanie zweryfikować wiele jego słynnych teorii. Korzystając z olbrzymiego zbioru danych, zawierającego zapisane sny, możemy łatwo zbadać, jak często pojawiają się w nich przedmioty o kształcie fallicznym. Żywność może być w tej sferze wdzięcznym przedmiotem zainteresowania uczonych, często występuje bowiem w snach, a to, co jemy, nierzadko ma kształt falliczny – banany, ogórki, hot dogi i tak dalej. Następnie możemy przyjrzeć się czynnikom, które sprawiają, że częściej śnimy o takich, a nie innych daniach, owocach czy warzywach – a zatem da się sprawdzić, jak często je spożywamy, czy ludzie w większości uważają je za smaczne, no i, owszem, czy rzeczywiście mają falliczną formę.

Możemy na przykład zbadać, czy dwie równie popularne potrawy, z których jedna kształtem przypomina męski narząd płciowy, pojawiają się w naszych snach z taką samą czy różną częstotliwością. Jeżeli okaże się, że potrawy o kształcie członka w erekcji nie śnią się ludziom częściej niż inne, to można będzie powiedzieć, że symbole falliczne nie odgrywają znaczącej roli w snach. Dzięki big data ta część teorii Freuda może więc naprawdę zostać sfalsyfikowana.

Informacje potrzebne mi do pracy zaczerpnąłem tym razem z Shadow – aplikacji, która prosi użytkowników, by zapisywali swoje sny – po czym odpowiednio zakodowałem potrawy wymienione w dziesiątkach tysięcy snów.

Co, najogólniej rzecz biorąc, sprawia, że śni nam się jedzenie? W głównej mierze decyduje o tym to, jak często spożywamy określone dania, owoce czy warzywa. Substancja, która śni się nam najczęściej, to woda, a wśród dwudziestu potraw najpowszechniej występujących w snach znajdują się kurczaki, chleb, kanapki i ryż – symbole ewidentnie niefreudowskie.

Drugim czynnikiem prognostycznym, wskazującym, jak często dana potrawa pojawi się w naszych snach, jest jej smakowitość. Dwa artykuły

spożywcze, które śnią się ludziom najczęściej i są wyraźnie niefreudowskie, lecz uchodzą powszechnie za smakowite, to czekolada i pizza.

No a co z potrawami o kształtach fallicznych? Czy przenikają do naszych snów nadspodziewanie często?

Nie.

Banany zajmują drugie miejsce, gdy chodzi o częstotliwość występowania w snach, plasują się jednak też na drugim miejscu wśród najczęściej spożywanych owoców. Ogórki występują na miejscu siódmym wśród najczęściej obecnych w naszych snach warzyw – ale znajdziemy je także na siódmym miejscu wśród najczęściej spożywanych warzyw. A zatem kształt ogórków nie jest potrzebny do tego, żeby wyjaśnić ich pojawianie się w naszych śniących umysłach. Znacznie rzadziej śnimy też o hot dogach niż o hamburgerach – których zjadamy daleko więcej niż hot dogów.

Ogólnie rzecz biorąc, dzięki zastosowaniu analizy regresji (metodzie pozwalającej socjologom rozdzielić wpływy różnorakich czynników) ustaliłem dla wszystkich owoców i warzyw, że falliczny kształt nie czyni występowania podłużnych płodów ziemi w naszych snach bardziej prawdopodobnym, niż wskazywałaby na to ich popularność. Ta teoria Freuda jest falsyfikowalna – i przynajmniej z punktu widzenia mojej analizy, fałszywa.

Weźmy teraz z kolei pomyłki freudowskie. Wiedeński psychoanalityk przypuszczał, że kiksy w mowie lub piśmie ujawniają nasze podświadome pragnienia, nierzadko seksualne. Czy możemy wykorzystać big data, by sprawdzić tę hipotezę? Oto jeden sposób: przekonajmy się, czy tego rodzaju błędy zdradzają „niegrzeczne" tendencje. Bo skoro ludzkie utajone skłonności seksualne wychodzą na jaw właśnie omyłkowo, to duża liczba pomyłek powinna zawierać takie słowa jak „penis", „kutas" czy „seks".

Przestudiowałem zatem w celach badawczych zbiór danych złożony z ponad czterdziestu tysięcy literówek odnotowanych przez pracowników Microsoftu[2]. Zbiór obejmował też błędy, które ludzie poprawiają od razu w trakcie pisania. W tak pokaźnej liczbie literówek

znalazło się wiele pomyłek natury seksualnej, takich jak wspomniany wyżej „penistrian". Ktoś napisał między innymi „sexurity" zamiast „security" [„bezpieczeństwo"] i „cocks" [„kutasy"] zamiast „rocks" [„skały", „kamienie"]. Znalazłem też jednak mnóstwo pomyłek niewinnych. Ludzie pisali „pindows" zamiast „windows" [„okna"], „fegetables" zamiast „vegetables" [„warzywa"], „aftermoons" zamiast „afternoons" [„popołudnia"], i „refriderators" zamiast „refrigerators" [„lodówki"].

Czy zatem liczba pomyłek seksualnych wśród literówek rzeczywiście okazała się nieprzeciętnie wysoka?

Żeby to zbadać, użyłem najpierw wspomnianego zbioru Microsoftu, chcąc sprawdzić, jak często omyłkowo przestawiamy konkretne litery. Obliczyłem na przykład, jak często zastępujemy „t" literą „s", a „g" literą „h". Potem stworzyłem program komputerowy, który popełniał takie błędy jak człowiek. Nazwijmy go Error Bot. Z taką samą częstotliwością, jak internauci w badaniach Microsoftu, zastępował „t" literą „s", a „g" literą „h". I tak dalej. Później przepuściłem przez ten program wszystkie słowa, w których, jak wynikało z badań Microsoftu, popełniali błędy ludzie. Mój bot próbował zatem pisać poprawnie „pedestrian", „rocks", „windows" i „refrigerator" – ale zamieniał „r" na „t" równie często jak człowiek, pisząc na przykład „tocks". Tak często jak człowiek zamieniał też „r" na „c" i czasem pisał „cocks" [„kutasy"].

Czego zatem w końcu się dowiedzieliśmy, porównując działania Error Bota z istotami ludzkimi, które są po prostu nieuważne z natury? Przestawiając litery, program popełnił kilka milionów błędów, w tym wiele takich, które można by nazwać freudowskimi. Zamiast „seashell" [„muszla"] pisał „sexshell", zamiast „lipstick" [„szminka"] – „lipsdick" [„fiut do ust"], a zamiast „luckiest" [„mający największe szczęście"] – „fuckiest" [„najchętniej posuwany/posuwana", „najbardziej chętny/chętna na seks", „najbardziej pierdolony"], i popełnił jeszcze wiele podobnych błędów. Kluczowym spostrzeżeniem było natomiast to, że Error Bot, niewyposażony, rzecz jasna, w podświadomość, popełniał błędy, które można uważać za seksualne, równie często jak ludzie z krwi i kości. Pamiętając, że trzeba jeszcze przeprowadzić w tej sferze kolejne badania, co zawsze powtarzamy my, przedstawiciele

nauk społecznych, można powiedzieć, że pomyłki o charakterze seksualnym nie zdarzają się ludziom częściej niż jakiekolwiek inne przypadkowe błędy.

Wynikałoby z tego, że aby zdarzały się takie literówki jak „penistrian", „sexurity" czy „cocks", nie musi istnieć żaden związek pomiędzy nimi i sferą tabu czy jakąś teorią umysłu, zgodnie z którą ludzie ujawniają swoje sekretne pragnienia w nieświadomych pomyłkach. Można je wyjaśnić, wskazując wyłącznie na przeciętną, przypadkową częstotliwość błędów literowych. Człowiek często się myli. Jeśli przytrafia się nam dość duża liczba pomyłek, w końcu powiemy lub napiszemy „lipsdick", „fuckiest" czy „penistrian". Jeżeli małpa będzie odpowiednio długo stukać w klawiaturę, to w końcu wystuka zdanie „być albo nie być", a człowiek w pewnym momencie napisze „penistrian".

Hipoteza Freuda, że pomyłki są odzwierciedleniem naszych nieuświadomionych pragnień, jest zatem falsyfikowalna – i zgodnie z moją analizą danych, fałszywa.

Big data mówi nam, że banan to po prostu banan, a „penistrian" to jedynie niepoprawnie zapisany „pedestrian".

Czy zatem Freud mylił się całkowicie we wszystkich swoich teoriach? Niezupełnie. Gdy po raz pierwszy uzyskałem dostęp do danych z PornHubu, znalazłem tam fascynujące informacje, które uderzyły mnie jako przynajmniej po trosze freudowskie. Można nawet powiedzieć, że jest to jedno z najbardziej zdumiewających spostrzeżeń, które poczyniłem w swojej pracy nad analizą danych: otóż szokująco wysoka liczba ludzi odwiedzających mainstreamowe portale pornograficzne poszukuje treści wizualnych przedstawiających stosunki kazirodcze.

Szesnaście na sto najpopularniejszych wyszukiwań w PornHubie, jednym z najchętniej odwiedzanych portali porno, dotyczy kazirodczych materiałów wideo. Muszę was uprzedzić, że są one dosyć obrazowe: „brat z siostrą", „macocha rucha się z synem", „matka z synem", „matka rucha się z synem" czy „prawdziwi brat i siostra".

Szukający kazirodczych scen mężczyźni najczęściej życzą sobie oglądać matki z synami. A kobiety? Na sto najpopularniejszych kobiecych wyszukiwań w PornHubie związek z kazirodztwem ma dziewięć i obejmują one podobne treści, tyle tylko, że zmienia się płeć rodziców i dzieci. To znaczy, że większość kazirodczych filmów, które chcą oglądać panie, przedstawia ojców i córki.

Nietrudno zauważyć w tych danych co najmniej ciche echo freudowskiego kompleksu Edypa. Freud sądził, że w dzieciństwie prawie wszyscy odczuwamy tego rodzaju pragnienie, które zostaje z upływem czasu stłumione – pragnienie kontaktów seksualnych z rodzicem płci przeciwnej. Szkoda, że wiedeński uczony nie żył na tyle długo, aby mógł zastosować swoje umiejętności analityczne do danych z PornHubu, gdzie zainteresowanie rodzicami płci przeciwnej zdradzają dorośli, i to bardzo wyraźnie, a prawie żadne z naszych pragnień stłumione nie jest.

Oczywiście dane z PornHubu nie powiedzą nam z całą pewnością, o kim naprawdę fantazjują internauci, kiedy oglądają takie filmy. Czy rzeczywiście wyobrażają sobie, że kochają się z własnymi rodzicami? Ale wyszukiwania w Google'u dostarczają kolejnych dowodów na to, że na świecie żyje mnóstwo osób, które mają podobne fascynacje.

Weźmy pod uwagę wszystkie sieciowe wyszukiwania w postaci „Chcę uprawiać seks z moją/moim…"[3]. Najczęściej frazę tę kończy słowo „mamą/tatą", a osiemdziesiąt dwa i siedem dziesiątych procent tak sformułowanych wyszukiwań dotyczy kazirodztwa. Nie chodzi jednak wyłącznie o taką formę wpisu. Na przykład wyszukiwania w postaci „odczuwam pociąg do…" dotyczą pragnień kazirodczych jeszcze częściej. Ryzykując, że rozczaruję Herr Freuda, spieszę przyznać, że treści tego typu nie są szczególnie liczne: w Stanach Zjednoczonych zaledwie kilka tysięcy osób rocznie wyznaje, że odczuwa pociąg do własnych matek. Ktoś powinien też poinformować Freuda, że wyszukiwania w Google'u czasami generalnie kierują się ku sferom tabu, chociaż niekoniecznie seksualnym, o czym będzie jeszcze mowa dalej.

Mimo to… Istnieje mnóstwo innych obiektów niestosownego zainteresowania, które, jak sądziłem, powinny się częściej pojawiać

w wyszukiwarce. Szef/szefowa? Podwładny/podwładna? Uczeń/uczennica? Terapeuta/terapeutka? Pacjent/pacjentka? Najlepsza przyjaciółka żony? Najlepsza koleżanka córki? Siostra żony? Żona najlepszego przyjaciela? Żadna z tych osób, których, do czego internauci się przyznają, niekiedy pożądamy, nie może jednak konkurować z matką. Być może w połączeniu z danymi z PornHubu to naprawdę coś znaczy.

Tym bardziej że generalizujące stwierdzenie Freuda, iż seksualność człowieka jest kształtowana przez doświadczenia z dzieciństwa, wspierają jeszcze inne dane z Google'a czy PornHubu, z których wynika, że przynajmniej panowie przechowują w pamięci niezwykle dużo fantazji erotycznych powiązanych z dzieciństwem. Jak pokazują wpisy żon na temat mężów, do najpopularniejszych marzeń dorosłych mężczyzn należy pragnienie, żeby nosić pieluszki i żeby żony karmiły ich piersią – to ostatnie życzenie, jak wspominałem wcześniej, szczególnie często występuje w Indiach. Dużym zainteresowaniem cieszą się też animowane filmy pornograficzne[4], w których w niecenzuralnych scenach występują postaci ulubione przez nastoletnich chłopców. Weźmy także pod uwagę najczęściej wyszukiwane przez mężczyzn zawody kobiece na portalach pornograficznych – mężczyźni pomiędzy osiemnastym i dwudziestym czwartym rokiem życia zazwyczaj chcą oglądać w akcji opiekunki do dzieci[5]. Podobnie panowie w przedziale wiekowym od dwudziestu pięciu do sześćdziesięciu czterech lat. A także sześćdziesięciopięcioletni i starsi. W pierwszej czwórce dla wszystkich tych grup wiekowych znajdują się jeszcze nauczycielki i czirliderki. Najwyraźniej początkowe lata życia rzeczywiście mają przemożny wpływ na fantazje dorosłych mężczyzn.

Nie miałem jak dotąd możliwości wykorzystania wszystkich tych bezprecedensowych informacji na temat seksualności, żeby dokładniej zbadać, w jaki sposób rodzą się nasze preferencje erotyczne. Ale w ciągu kilku najbliższych dziesięcioleci zarówno ja, jak i inni badacze z zakresu nauk społecznych z pewnością zdołamy stworzyć nowe, falsyfikowalne teorie dotyczące życia seksualnego dojrzałego człowieka i zweryfikować je, konfrontując z aktualnym zbiorem danych na ten temat.

Już dzisiaj mogę przewidzieć parę podstawowych wątków, które niewątpliwie staną się częścią teorii seksualności bazującej na analizie danych. Oczywiście nie będzie ona identyczna z tą, którą zaproponował Freud, opisujący konkretne, jasno określone etapy erotyzmu dziecięcego i nasze tłumione pragnienia. Opierając się jednak na moim pierwszym kontakcie z danymi z PornHubu, jestem absolutnie pewien, że ostateczny werdykt w procesie życia płciowego osób dorosłych będzie odnosił się do niektórych kluczowych tez postawionych przez Freuda, a dzieciństwo odegra w nim istotną rolę. Ważne miejsce zajmą również matki.

Poddanie teorii Freuda takiej analizie najprawdopodobniej byłoby niemożliwe jeszcze dziesięć, a już na pewno osiemdziesiąt lat temu, czyli wciąż za życia wiedeńskiego uczonego. Zastanówmy się więc, dlaczego te źródła danych są nam pomocne. Ćwiczenie to pozwoli zrozumieć powody, czyniące big data rzeczywiście potężnym narzędziem badawczym.

Jak pamiętacie, powiedzieliśmy, że sam olbrzymi zbiór danych nie będzie automatycznie i samoistnie generował takich czy innych odkryć. Wielkość zbioru danych sama w sobie jest często przeceniana. Z czego wynika zatem moc big data? I dlaczego wywoła rewolucję, która sprawi, że wkrótce spojrzymy na siebie zupełnie inaczej? Według mnie big data odznacza się czterema unikatowymi mocami, a analiza teorii Freuda dobrze je ilustruje.

Skonstatujmy na początek, że jak pewnie zauważyliście, w rozważaniach o Freudzie traktujemy pornografię serio. I z danych na temat pornografii często będziemy w niniejszej książce korzystać. Zabrzmi to może nieco zaskakująco, ale socjologowie rzadko sięgają po materiały źródłowe ze świata pornografii, zadowalając się w większości tradycyjnymi zbiorami informacji, pochodzących z ankiet i sondaży, na których opierali się podczas swojej dotychczasowej kariery zawodowej. Wystarczy jednak chwila refleksji, aby się zorientować, że powszechne zainteresowanie pornografią, jak również towarzyszące jej dane w postaci wpisów w sieci i oglądanych przez internautów treści wizualnych, stanowią najważniejsze elementy na drodze do

zrozumienia ludzkiej seksualności od czasów… No, właściwie chyba w całej naszej historii. Schopenhauer, Nietzsche, Freud i Foucault dostaliby po prostu ślinotoku na widok tych zbiorów danych, które za ich życia jeszcze nie istniały. Nie istniały też nawet dwadzieścia lat temu – ale istnieją dzisiaj. Obecnie dysponujemy wieloma wyjątkowymi źródłami informacji, dotyczących całego szeregu tematów i dających nam wgląd w takie sfery życia, które wcześniej rozświetlaliśmy jedynie domysłami. Nowe typy źródeł danych to pierwsza moc big data.

Dane z obszaru pornografii i z wyszukiwarki Google'a są nie tylko nowe, ale też prawdziwe w tym sensie, że szczere, nieskłamane. W epoce precyfrowej nie ujawnialiśmy swoich wstydliwych myśli przed innymi ludźmi. Dziś również się ich wstydzimy, ale nie taimy ich przed Internetem i niektórymi wyszukiwarkami, takimi jak Google czy PornHub, które chronią naszą anonimowość. Te portale funkcjonują jako coś w rodzaju cyfrowego serum prawdy – dlatego jesteśmy na przykład w stanie pokazać, jak rozpowszechnione jest zainteresowanie kazirodztwem. Big data pozwala nam także przekonać się, czego ludzie naprawdę chcą i co naprawdę robią, a nie co mówią, że chcą, i nie co mówią, że robią. Gromadzenie prawdziwych danych to druga moc big data.

Ponieważ dysponujemy teraz tak wielkim zbiorem danych, otrzymujemy ważne informacje na temat nawet bardzo skromnych ułamków populacji. Możemy na przykład porównać liczbę osób, którym śnią się ogórki, z liczbą osób, którym śnią się raczej pomidory. Możliwość dokładnego przyjrzenia się małym grupom społecznym to trzecia moc big data.

W trakcie swojej ekspresowej analizy teorii Freuda nie odwołałem się do kolejnej imponującej mocy big data, ale może zrobię to w przyszłości – i przeprowadzę kilka szybkich, kontrolowanych eksperymentów, dzięki którym bada się przyczynowość, nie tylko korelacje. Tego rodzaju doświadczenia zlecają dzisiaj w większości jedynie firmy i przedsiębiorstwa, wkrótce jednak takie eksperymenty staną się potężnym narzędziem również dla przedstawicieli nauk społecznych. Możliwość przeprowadzania licznych, a przy tym szybkich badań to czwarta moc big data.

Nadeszła zatem pora, by przyjrzeć się każdej z owych mocy po kolei i bardziej szczegółowo odpowiedzieć na pytanie, dlaczego nie należy lekceważyć wielkich zbiorów danych.

ROZDZIAŁ TRZECI
DANE PRZEOBRAŻONE

Ulice prawie całego Manhattanu wyglądają o godzinie szóstej rano na niemal całkowicie opustoszałe. Zamknięte sklepy są na dole od frontu osłonięte stalowymi żaluzjami antywłamaniowymi, a mieszkania na wyższych kondygnacjach ciemne i milczące.

Ale w jeden piątek każdego miesiąca na piętrach zajmowanych przez ogromny bank inwestycyjny Goldman Sachs na dolnym Manhattanie nawet wtedy świeci się jaskrawe światło, a windy przewożą tysiące pracowników, zdążających na swoje miejsca przy biurkach. O siódmej rano większość z nich zostanie już zajęta.

Nie popełnilibyśmy nietaktu, mówiąc o tej części miasta każdego innego dnia, że jest senna, ale jednego piątkowego poranka w miesiącu na Manhattanie wyczuwa się napięcie i poruszenie. Bo tego dnia dociera tu pewna wiadomość, która za każdym razem trzęsie giełdą papierów wartościowych.

Zaledwie kilka minut po jej ujawnieniu zostanie przekazana przez portale informacyjne, a najwyżej kilka sekund później stanie się przedmiotem dyskusji oraz sporów i będzie głośno rozbierana na czynniki pierwsze zarówno w Goldman Sachs, jak i setkach innych firm finansowych. Lecz w dzisiejszych czasach to, co naprawdę liczy się w bankowości, rozgrywa się w ciągu kilku milisekund, nie sekund. Goldman oraz inne przedsiębiorstwa finansowe zapłaciły dziesiątki milionów dolarów za kable światłowodowe, by skrócić czas przesyłu informacji z Chicago do New Jersey o zaledwie cztery milisekundy (z siedemnastu do trzynastu). Firmy finansowe dysponują specjalnymi algorytmami[1] do odczytu danych i prowadzenia opierających się na nich interesów – zajmuje im to właśnie kilka milisekund. Dlatego po

ujawnieniu owej niezwykle ważnej wiadomości zmiany na rynku następują szybciej, niż trwa mgnienie oka.

A czego dotyczy niezwykle ważna wiadomość, warta tyle pieniędzy dla Goldmana i wielu innych instytucji finansowych?

Miesięcznej stopy bezrobocia w kraju.

Ta informacja, która ma tak poważny wpływ na giełdę, że instytucje finansowe robią, co tylko mogą, aby jak najszybciej ją uzyskać, przeanalizować i rozpocząć na jej podstawie odpowiednie działania, jest jednak wynikiem telefonicznego sondażu, prowadzonego przez Bureau of Labor Statistics [Biuro Analizy Statystycznej Rynku Pracy], i kiedy zostaje opublikowana, liczy sobie już całe trzy tygodnie – czyli dwa miliardy milisekund.

Zdziwicie się zapewne bardziej niż trochę, że z jednej strony firmy wydają miliony dolarów na to, aby przyspieszyć przekaz tych danych o chociażby mikroskopijne wartości, a z drugiej rząd federalny oblicza stopę bezrobocia tak powoli.

Kiedy w 2011 roku Alan Krueger, współpracownik prezydenta Obamy i szef jego Council of Economic Advisors [Rady Ekonomicznej], obejmował swoje stanowisko, starał się przyspieszyć publikację tych niezwykle ważnych informacji. Bezskutecznie. „Albo BLS nie ma na to środków – powiedział – albo jego pracownicy utknęli umysłowo w XX wieku”[2].

Rząd najwyraźniej nie zamierza w najbliższej przyszłości przyspieszyć w tej kwestii, ale czy istnieje jakiś sposób, by nieco szybciej chociaż z grubsza oszacować miesięczną stopę bezrobocia? Czy w czasach nowoczesnych technologii, kiedy niemal każde kliknięcie w Internecie zostaje gdzieś zapisane, naprawdę musimy czekać całymi tygodniami, żeby się dowiedzieć, ile osób w kraju pozostaje bez pracy?

Jedno potencjalne rozwiązanie zaproponował były pracownik Google’a Jeremy Ginsberg, który zauważył, że rząd federalny wypuszcza informacje dotyczące zdrowia obywateli z dużym opóźnieniem, dokładnie tak samo jak dane o bezrobociu. Centers for Disease Control and Prevention [Ośrodki Kontroli Chorób i Prewencji Zdrowotnej] publikują na przykład statystyki zachorowań na grypę z tygodniowym

poślizgiem, mimo że szpitale i lekarze bardzo by skorzystali na znacznie szybszym dostępie do tych danych.

Ginsberg przypuszczał, że osoby chore na grypę szukają w sieci odpowiedzi na pytania związane z tą chorobą – czyli że, krótko mówiąc, zwierzają się z jej symptomów wyszukiwarce – i że dzięki temu można ustalić stosunkowo dokładną liczbę zakażeń. Miał rację, bo wyszukiwania takie jak „objawy grypy" czy „bóle mięśniowe" szybko okazały się ważnymi wskaźnikami[3] tempa rozprzestrzeniania się choroby[4*].

Tymczasem informatycy z Google'a stworzyli serwis internetowy, Google Correlate, który umożliwia badaczom spoza firmy prowadzenie tego rodzaju analiz w bardzo wielu sferach naszego życia, nie tylko w zakresie opieki zdrowotnej. Dzisiaj możemy więc wyodrębnić interesujące nas dane, które Google rejestruje w czasie, i sprawdzić, jakie wyszukiwania występują najczęściej w korelacji z ich zbiorem.

Na przykład korzystając z Google Correlate, Hal Varian, główny ekonomista Google'a, i ja ustaliliśmy, jakie wpisy internetowe pozwalają przewidywać zmiany cen nieruchomości[4]. Gdy ceny rosną, Amerykanie na ogół wrzucają w wyszukiwarki frazy typu „hipoteka 80/20", „budowa nowych domów" czy „stopa aprecjacji", a kiedy spadają, pojawia się „proces *short sale*", „hipoteka o straconej wartości" albo „odsetki hipoteczne wolne od podatku".

Czy zatem wpisy internetowe mogą być takim papierkiem lakmusowym bezrobocia jak w wypadku zachorowań na grypę albo cen nieruchomości? Czy dzięki temu, czego ludzie poszukują w Google'u, możemy stwierdzić, ile osób pozostaje bez pracy, i czy da się to zrobić dużo wcześniej, zanim rząd ogłosi wyniki swojego sondażu?

Pewnego dnia przepuściłem amerykańską stopę bezrobocia w latach 2004–2011 przez Google Correlate.

Jak myślicie, jaka fraza z trylionów wpisów z tego czasu pojawiała się w związku z bezrobociem najczęściej? Sądzicie może, że „pośredniak" albo coś w tym rodzaju? Owszem, plasował się wysoko, lecz nie na samej górze. „Nowe miejsca pracy"? One również figurowały wysoko, ale nie na szczycie.

W okresie, który objęła moja analiza, najczęściej występującym wpisem – choć jego forma się zmienia – była nazwa „Slutload". Tak jest, ludzie najchętniej wyszukiwali portal pornograficzny. Na pierwszy rzut oka i pierwszy rumieniec może się to wydawać dosyć dziwne, należy jednak przypuszczać, że osoby bezrobotne mają sporo wolnego czasu. Wiele z nich siedzi samotnie w domu i się nudzi. Inna często wpisywana fraza – tym razem poza nieprzyzwoitymi portalami – brzmiała: „pasjans pająk". To chyba również nie dziwi, bo nadal mówimy o ludziach, którzy najprawdopodobniej mają dużo czasu.

Nie twierdzę oczywiście, opierając się na jednej analizie, że śledzenie takich wyszukiwań jak „Slutload" czy „pasjans pająk" to najlepszy sposób, aby określić stopę bezrobocia. Rozrywki, jakim się oddają bezrobotni, zmieniają się w czasie (w pewnym okresie najpopularniejszym wpisem–korelatem bezrobocia był „rawtube", inny portal porno), a żadna z nich sama z siebie ani trochę nie oddaje różnorodności w grupie bezrobotnych. Ogólnie rzecz biorąc, przekonałem się jednak, że cała mieszanka wpisów związanych z rozrywką może pomóc w ustaleniu stopy bezrobocia – i powinna być częścią skutecznego modelu predykcyjnego.

Powyższy przykład ilustruje pierwszą moc big data, każąc nam zweryfikować wyobrażenia o tym, co traktujemy jako dane. Wartość big data często wcale nie kryje się w wielkości zbioru, tylko w tym, że może nam zaproponować nowe obszary informacji, którym warto się przyjrzeć – takich informacji, jakich nigdy przedtem nie zbierano.

Przed wyszukiwarką Google'a mieliśmy do dyspozycji dane na temat niektórych sposobów spędzania wolnego czasu – znaliśmy na przykład liczbę biletów sprzedawanych na seanse kinowe – które mogły podpowiedzieć badaczom, ile Amerykanie rzeczywiście go mają. Ale dopiero dzisiaj możemy sprawdzić, jak często internauci stawiają w sieci pasjanse i ile czasu poświęcają na oglądanie pornografii, a są to informacje zarówno nowe, jak i niezwykle znaczące. W tym konkretnym wypadku analiza danych może nam pomóc określić znacznie prędzej, jak radzi sobie krajowa gospodarka – w każdym razie dopóki rząd nie zacznie szybciej przeprowadzać sondaży i upowszechniać ich wyników.

Życie na kampusie Google'a w Mountain View w Kalifornii radykalnie różni się od pracy w głównej siedzibie banku Goldman Sachs na Manhattanie. Biura Google'a są właściwie puste jeszcze o dziewiątej rano. Jeśli kręcą się w nich jacyś pracownicy, to zapewne po to, by zjeść darmowe śniadanie, na które składają się naleśniki bananowo-jagodowe, jajecznica z samych białek i filtrowana woda z ogórków. Niektórzy przebywają chyba poza miastem: są na jakimś spotkaniu w Boulder czy Las Vegas albo na darmowej wycieczce narciarskiej nad jeziorem Tahoe. Mniej więcej w porze lunchu zapełnią się boiska do gry w siatkówkę plażową i trawiaste boisko do piłki nożnej. Najlepsze burrito w życiu jadłem w meksykańskiej restauracji Google'a.

Jak to możliwe, że jedna z największych i najbardziej konkurencyjnych firm technologicznych na świecie pracuje na takim luzie i jest dla swoich pracowników tak hojna? Google okiełznał big data jak żadna inna firma, przekształcając zbiór danych w maszynkę do robienia pieniędzy. Wspominam ją często w tej książce, bo to właśnie ta wyszukiwarka stanowi zdecydowanie najważniejsze źródło big data. Trzeba jednak pamiętać, że sukces Google'a sam został zbudowany na zbiorze nowych danych.

Jeżeli jesteście w takim wieku, że korzystaliście z Internetu jeszcze w XX stuleciu, to może pamiętacie pierwsze wyszukiwarki z tamtych lat – MetaCrawler, Lycos czy AltaVista, żeby wymienić choć kilka. I chyba też pamiętacie, że bywały one dość zawodne. Jeśli sprzyjało wam szczęście, czasami udawało im się znaleźć to, czego potrzebowaliście, ale rzadko. Jeszcze w drugiej połowie lat dziewięćdziesiątych po wpisaniu w największą ówczesną wyszukiwarkę hasła „Bill Clinton" pośród najczęstszych wyników pojawiała się na przykład jakaś przypadkowa strona, głosząca po prostu, że „Bill Clinton jest do dupy", albo witryna, gdzie można było znaleźć marnej jakości dowcipy o Clintonie – czyli że dostawaliśmy z tych źródeł stosunkowo mało istotne informacje o ówczesnym prezydencie Stanów Zjednoczonych.

W 1998 roku na rynku pojawił się Google, którego wyniki wyszukiwania okazały się bezdyskusyjnie lepsze od całej konkurencji. Jeżeli w 1998 roku wpisaliście w Google'a „Bill Clinton", dostawaliście

w odpowiedzi adres jego strony internetowej, adres mejlowy Białego Domu i najlepsze biogramy prezydenta, jakie można było wtedy znaleźć w sieci. Google czynił czary.

Co takiego zrobili jego założyciele, Sergey Brin i Larry Page, że ich wyszukiwarka działała inaczej?[5].

Konkurencyjne algorytmy znajdowały dla swoich użytkowników strony, na których najczęściej pojawiała się szukana fraza. Jeżeli interesowały cię informacje o Billu Clintonie, to znajdowały w Internecie te strony, na których go najczęściej wzmiankowano. Istniało wiele powodów, dla których taki system był niedoskonały, a jednym z nich jest to, że łatwo go było oszukać. Na przykład portal z dowcipami, gdzie na jakiejś stronie ukrywałby się, powiedzmy, tekst typu „Bill Clinton Bill Clinton Bill Clinton Bill Clinton", zostałby sklasyfikowany wyżej niż oficjalna strona Białego Domu[5*].

Brin i Page znaleźli natomiast sposób rejestrowania informacji nowego typu, i to o wiele bardziej wartościowych niż prosta liczba powtórzeń danego wyrazu. Portale internetowe często tworzyły związane z poruszanym tematem linki do stron, które ich zdaniem mogły być pomocne w jego przybliżeniu. Na przykład jeśli Billa Clintona wspominał „New York Times", odsyłał czytelników, którzy kliknęli jego nazwisko, na oficjalną stronę Białego Domu.

Każdy portal tworzący taki link w pewnym sensie wyrażał opinię o tym, gdzie można znaleźć najlepsze wiadomości o Clintonie. Brinowi i Page'owi udało się zebrać wszystkie tego rodzaju opinie na każdy temat. Ściągali je[6] z „New York Timesa", od milionów użytkowników systemu obsługi list dyskusyjnych Listserv, setek blogerów i innych internautów. Jeżeli cała masa ludzi uważa, że najważniejszym linkiem związanym z hasłem „Bill Clinton" jest jego oficjalna strona internetowa, to można przypuszczać, że większość internautów, wpisująca w wyszukiwarkę nazwisko prezydenta, chciałaby zobaczyć właśnie tę stronę.

Te linki niesłychanie ułatwiały dostęp do najbardziej użytecznych informacji na żądany temat i były danymi, których inne wyszukiwarki nie brały nawet pod uwagę. Trzeba tu podkreślić, że Google nie

zdominował rynku wyszukiwarek wyłącznie dlatego, że zgromadził więcej danych niż jego konkurenci, tylko dlatego, że miał lepsze dane. W ciągu niecałych dwóch lat napędzany analizą linków Google stał się najpopularniejszą wyszukiwarką internetową na świecie, a Brin i Page są dzisiaj warci ponad sześćdziesiąt miliardów dolarów.

To, co jest prawdą o Google'u, obchodzić musi wszystkich, którzy wykorzystują analizę danych, aby zrozumieć świat. Rewolucja big data polega nie tyle na gromadzeniu coraz większej ilości informacji, ile na gromadzeniu informacji dobrych, odpowiednich do realizacji zadanego celu.

Internet to wszakże nie jedyne miejsce, gdzie można zbierać nowe informacje i gdzie dostęp do odpowiednich danych może przynieść przełomowe rezultaty. Niniejsza książka dotyczy w dużym stopniu tego, w jaki sposób materiały z sieci mogą nam pomóc lepiej poznać człowieka, ale następny podrozdział nie ma nic wspólnego z danymi z Internetu. Ba, nie ma nawet nic wspólnego z ludźmi, dobrze ilustruje jednak główną tezę niniejszego rozdziału, czyli nieocenioną wartość nowych, niekonwencjonalnych danych. Zasady, których możemy się z niego nauczyć, pozwolą nam lepiej zrozumieć zachodzącą właśnie na naszych oczach cyfrową rewolucję informacyjną.

CIAŁO JAKO ZBIÓR DANYCH

Latem 2013 roku pewną niewielką stajnię w północnej części stanu Nowy Jork zajmował więcej niż średniej wielkości, kasztanowaty koń o czarnej grzywie. Był jednym ze stu pięćdziesięciu dwóch jednorocznia ków wystawionych na sprzedaż w odbywającej się w sierpniu aukcji Fasig-Tipton Selected Yearling Sale w Saratoga Springs i jednym z dziesięciu tysięcy, które miały trafić na licytację tamtego roku.

Kiedy zamożni mężczyźni albo bogate kobiety wykładają duże pieniądze na konia wyścigowego, to chcą mieć zaszczyt wybrania dla niego imienia. Dlatego ten kasztanek jeszcze go nie miał i jak większość zwierząt na wyprzedaży nosił tylko numer swojej stajni: 85.

Niewiele wyróżniało tego konia wśród innych. Miał dobry, chociaż wcale nie rewelacyjny rodowód. Jego ojciec, Pioneerof (sic!) the Nile, był znakomitym koniem wyścigowym, lecz jego inne dzieci nie odnosiły zbyt wielkich sukcesów na torze. Wątpliwości budził także stan zdrowia Numeru 85. Miał na przykład zadrapanie w okolicach pęciny, niektórzy kupujący obawiali się więc, że jest to oznaka jakiejś kontuzji.

Ówczesnym właścicielem tego konia był egipski magnat piwny Ahmed Zayat, który przyjechał na aukcję z nadzieją na jego sprzedaż i kupno kilku innych jednoroczniaków.

Jak prawie wszyscy właściciele, Zayat wynajął zespół ekspertów, którzy mieli mu dopomóc w wyborze koni przy kupnie. Specjaliści Zayata różnili się jednak nieco od tych, z których usług korzystali prawie wszyscy nabywcy poza nim. Typowy koński znawca, jakiego można spotkać na takich imprezach, jest zwykle w średnim wieku, pochodzi z Kentucky albo florydzkiej wsi i nie ma szczególnie imponującego wykształcenia, choć jego rodzina zazwyczaj już od pokoleń zajmuje się handlem końmi. Eksperci Egipcjanina wywodzili się jednak z niewielkiej firmy o nazwie EQB, a jej szef nie należał do oldskulowych znawców koni. Nazywał się Jeff Seder, był urodzonym w Filadelfii ekscentrykiem i mógł się pochwalić kilkoma dyplomami z Harvardu.

Zayat ludzi z EQB zatrudniał już wcześniej, znał więc sposób ich pracy. Po kilku dniach, bo tyle zajmowała im ocena zwierząt, mieli do niego wrócić z mniej więcej pięcioma nowymi propozycjami na miejsce Numeru 85.

Tym razem stało się inaczej. Zespół Sedera po powrocie stwierdził, że nie może spełnić jego prośby. Po prostu nie mogli mu zarekomendować żadnego ze stu pięćdziesięciu jeden pozostałych koni wystawianych tego dnia na sprzedaż. Zgłosili natomiast nieoczekiwaną i niemal desperacką prośbę. Zayat absolutnie i z całą pewnością nie powinien sprzedawać Numeru 85. Bo ten koń, oświadczyli specjaliści z EQB, nie tylko jest najlepszym koniem na aukcji, ale też najlepszym koniem roku, a może nawet dekady. „Niech pan lepiej sprzeda dom – zaklinali Zayata. – Byle nie tego konia”[8].

Następnego dnia bez zbędnych fanfar konia o numerze 85 kupił za trzysta tysięcy dolarów człowiek nazwiskiem Incardo Bloodstock. Jak później ujawniono, był to pseudonim Ahmeda Zayata. W odpowiedzi na prośby Sedera odkupił własnego konia, była to zatem sprawa niemal bez precedensu (zasady aukcyjne nie pozwalały Zayatowi po prostu wycofać konia ze sprzedaży, musiał więc przeprowadzić transakcję pod pseudonimem). Drożej niż Numer 85 sprzedano aż sześćdziesiąt dwa jednoroczniaki, przy czym dwa z nich osiągnęły ceny powyżej miliona dolarów.

Po trzech miesiącach Zayat wybrał w końcu imię dla Numeru 85: American Pharoah. Półtora roku później w ciepły sobotni wieczór na przedmieściach Nowego Jorku American Pharoah został pierwszym koniem od przeszło trzydziestu lat, który zdobył Potrójną Koronę.

Co takiego wiedział Jeff Seder o koniu numer 85, czego nie wiedział nikt inny? Jak to się stało, że facet po Harvardzie nauczył się tak dobrze oceniać sportową wartość koni?

Pierwszy raz spotkałem się z Sederem[9], wtedy sześćdziesięcioczterolatkiem, upalnego czerwcowego popołudnia w Ocala na Florydzie, ponad rok po zdobyciu Potrójnej Korony przez American Pharoah. Były to trwające tydzień widowiskowe zawody dla dwulatków, zakończone aukcją podobną do tej z 2013 roku, na której Zayat sam od siebie odkupił swojego ogiera.

Seder ma dźwięczny głos, przypominający trochę głos Mela Brooksa, gęste włosy i porusza się wyraźnie sprężystym krokiem. Był w szelkach, spodniach khaki i czarnej koszuli z logo swojej firmy. Nosił aparat słuchowy.

Przez następne trzy dni opowiadał mi historię swojego życia – i o tym, jak doszedł do tego, że tak dobrze nauczył się oceniać konie wyścigowe. Nie zmierzał prostą drogą. Ukończył na Harvardzie swoje pierwsze studia *magna cum laude*, należał do zrzeszającej najlepszych studentów organizacji Phi Beta Kappa, po czym zrobił jeszcze na tej samej uczelni drugi dyplom z prawa i biznesu. W wieku dwudziestu sześciu lat zaczął pracować jako analityk dla Citigroup w Nowym Jorku, gdzie jednak czuł się nieszczęśliwy i wypalony. Któregoś dnia, siedząc w atrium w nowych biurach firmy na Lexington Avenue, zaczął się

przyglądać dużemu muralowi, przedstawiającemu otwarte pole. Malowidło przypomniało mu o jego miłości do wsi i koni. Wrócił do domu i wciąż tkwiąc w trzyczęściowym garniturze, przejrzał się w lustrze. Zrozumiał, że nie nadaje się do pracy w bankowości i nie powinien mieszkać w Nowym Jorku. Rano złożył w firmie wymówienie.

Przeprowadził się na wieś, do Pensylwanii, i imał się tam wielu różnych zajęć, od handlu tekstyliami po medycynę sportową, aż wreszcie poświęcił się bez reszty swojej pasji, czyli zaczął oceniać sportowe szanse koni wyścigowych. Matematyka jest w tej dziedzinie brutalna. Z tysiąca dwulatków prezentowanych na aukcji w Ocala, jednej z najbardziej prestiżowych w kraju, może pięć wygra kiedyś jakąś gonitwę o liczącą się finansowo nagrodę. A pozostałych dziewięćset dziewięćdziesiąt pięć zwierząt? Mniej więcej jedna trzecia[10] okaże się zbyt wolna. Jedna trzecia dozna kontuzji, w większości dlatego, że ich nogi nie będą w stanie wytrzymać olbrzymiego wysiłku fizycznego, jakiego wymaga bieg ze wszystkich sił (każdego roku na amerykańskich torach zdychają setki koni[11], głównie właśnie z powodu złamanych nóg[12]). I wreszcie jedna trzecia będzie cierpieć na coś, co można by nazwać syndromem Bartleby'ego. Bartleby, skryba z genialnego opowiadania Hermana Melville'a, przestaje nagle pracować i na każdą prośbę ze strony chlebodawcy odpowiada, że wolałby jej więcej nie spełniać. A wiele koni już u progu swojej kariery sportowej najwyraźniej zdaje sobie sprawę, że wcale nie muszą biegać, jeśli nie mają na to ochoty. Zaczynają czasem gonitwę w szybkim tempie, lecz w pewnym momencie zwalniają lub w ogóle się zatrzymują. Bo po co biegać po jajowatym torze ile sił w nogach, zwłaszcza jeśli bolą cię kopyta i pęciny? Niektóre konie wyścigowe uznają zatem, że lepiej tego nie robić (mam słabość do wszystkich Bartlebych, zarówno wśród zwierząt, jak i ludzi).

Skoro szanse na sukces są tak nikłe, to jak można wskazać dobrego konia? Panowała rozpowszechniona opinia, że najłatwiej przewidzieć, czy koń będzie wygrywał, jeżeli przyjrzymy się jego rodowodowi, a być ekspertem od koni oznacza, że umie się wyrecytować wszystko, co można by chcieć wiedzieć o ich ojcach, matkach, dziadkach, babciach, braciach i siostrach. Fachowcy powiadają na przykład, że duży koń

„osiągnął swoje rozmiary prawowicie", jeśli w jego drzewie genealogicznym ze strony matki występuje wielu rosłych przodków.

Jest jednak pewien problem. Rodowód ma wprawdzie znaczenie, lecz za jego pomocą można jedynie cząstkowo wyjaśnić sukces konia. Weźmy pod lupę wyniki najbliższego rodzeństwa wszystkich zdobywców tytułu Konia Roku, najbardziej prestiżowej dorocznej nagrody na świecie. Te zwierzęta mogą się poszczycić najlepszym możliwym pochodzeniem, bo przecież ich drzewo genealogiczne jest takie samo jak koni, które przeszły do historii. Mimo to więcej niż trzy czwarte[13] z nich nie wygrywa w życiu ani jednej ważniejszej gonitwy. Analiza danych mówi nam wyraźnie, że ten tradycyjny sposób przewidywania przyszłych sukcesów konia na torze wyścigowym pozostawia wiele do życzenia.

Właściwie jednak nie powinniśmy się dziwić, że rodowód wcale nie jest takim dobrym materiałem predykcyjnym. Pomyślcie o ludziach. I wyobraźcie sobie właściciela drużyny NBA, który opierając się na rodowodach, kupuje do niej graczy, kiedy mają dopiero dziesięć lat. Wynajałby pewnie fachowca, by przyjrzał się Earvinowi Johnsonowi III, synowi „Magica" Johnsona. „Na razie ładnie rośnie – mógłby powiedzieć fachowiec. – Prawowity wzrost, z linii Johnsonów. Powinien mieć w przyszłości świetną orientację na boisku, wzrost i szybkość, i nie będzie chyba egoistą. Sprawia wrażenie towarzyskiego, świetna osobowość. Pewny krok. Łatwo nawiązuje kontakty. Warto na niego postawić". Niestety dwadzieścia dwa lata później właściciel tego zespołu dostałby gracza wzrostu sześciu stóp i dwóch cali (czyli niskiego jak na zawodowego koszykarza), a przy tym będącego blogerem modowym dla „E!". Earvin Johnson III[14] mógłby mu bardzo pomóc w projektowaniu strojów dla zawodników, lecz nie okazałby się szczególnie przydatny na parkiecie.

Właściciel drużyny z NBA, który kompletowałby zawodników, tak jak wybiera się konie wyścigowe, oprócz modowego blogera rzuciłby się też pewnie na Jeffreya i Marcusa, synów Michaela Jordana, którzy okazali się zaledwie średniej klasy graczami ligi akademickiej – możemy więc im tylko życzyć powodzenia w meczu przeciwko Cleveland Cavaliers z LeBronem Jamesem na czele, choć jego matka mierzy zaledwie pięć

stóp i pięć cali wzrostu[15]. Albo wyobraźcie sobie kraj, który na podstawie drzewa genealogicznego wybiera prezydenta: wtedy przewodziliby w nim ludzie tacy jak George W. Bush (sorki, nie mogłem się powstrzymać).

Znawcy koni korzystają też oczywiście z innych informacji niż rodowodowe. Analizują na przykład chody dwulatków i uważnie im się przyglądają. W Ocala prowadziłem jednak na tyle długie rozmowy z różnymi ekspertami, aby się zorientować, że w gruncie rzeczy nie zgadzają się ze sobą co do tego, czego szukają.

Do wszystkich tych mnożących się sprzeczności i niepewności dochodzą kupujący, którzy jak się zdaje, dysponują czasem niemal nieograniczonymi funduszami, co w sumie składa się na sporą liczbę pomyłek na końskim rynku. Dziesięć lat temu koń numer 153 był najszybszym dwulatkiem, wszystkim ekspertom wydawał się piękny, a do tego jako potomek pary Northern Dancer / Secretariat, dwójki najlepszych koni wyścigowych wszech czasów, miał wspaniały rodowód. Chcieli go kupić pewien irlandzki miliarder i szejk z Dubaju. Doszło nawet między nimi do wojny licytacyjnej, która szybko przerodziła się w rywalizację czysto ambicjonalną. Setki amatorów koni wyścigowych przyglądało się coraz wyższym stawkom, aż wreszcie dwulatka sprzedano za szesnaście milionów dolarów, zdecydowanie najwyższą kwotę w historii. Ale Numer 153, któremu nadano imię Green Monkey[16], pobiegł w życiu tylko w trzech wyścigach, zarobił zaledwie dziesięć tysięcy dolarów i został szybko odesłany na emeryturę.

Sedera nigdy nie interesowały tradycyjne kryteria oceny konia, lecz konkretne informacje. Zamierzał przebadać różne cechy koni wyścigowych i sprawdzić, które mają wpływ na ich wyniki. Trzeba tu zauważyć, że chociaż opracował swój program na pięć lat przed wynalezieniem Internetu, to jego strategia w znacznym stopniu opierała się na analizie danych. Wszystko, czego się przy okazji nauczył, będzie przydatne każdemu, kto korzysta z big data.

Przez długie lata poszukiwania Sedera dostarczały mu samych rozczarowań. Mierzył na przykład wielkość końskich chrap i stworzył pierwszą i największą na świecie bazę danych na ten temat, która, nawiasem mówiąc, w końcu przyniosła zyski. Okazało się jednak, że

wielkość chrap nie ma związku z dobrymi wynikami na torze. Robił też koniom EKG i badał odcięte nogi padłych zwierząt, by mierzyć wielkość ich mięśni szybkokurczliwych. Raz chwycił nawet szpadel przed wejściem do stajni, żeby ważyć końskie ekskrementy, ponieważ pojawiła się teoria, że koń będzie wolniejszy, jeżeli przed wyścigiem straci za dużo na wadze. Nie znalazł jednak czynników, które miałyby wpływ na sukcesy zwierząt na torze.

Potem, dwanaście lat temu, dokonał pierwszego istotnego odkrycia. Postanowił przeprowadzać pomiary wielkości wewnętrznych organów koni. Ówczesna technologia na to nie pozwalała, skonstruował więc własny przenośny ultrasonograf. Rezultaty były zdumiewające. Okazało się, że wielkość serca, a zwłaszcza jego lewej komory, jest niezwykle ważnym elementem, dzięki któremu można prognozować najważniejszą zmienną – wyniki zwierzęcia na torze. Innym istotnym dla sukcesów organem, odkrył Seder, jest śledziona: konie z małą śledzioną właściwie nigdy nie odnosiły zwycięstw.

Powiodło mu się jeszcze w dwóch kwestiach. Po pierwsze, zapisał cyfrowo nagrania galopujących koni i odkrył, że wpływ na dobre wyniki mają niektóre chody. Poza tym ustalił, że są dwulatki zaczynające rzęzić już po przebiegnięciu jednej ósmej mili. Ceny takich zwierząt sięgają niekiedy nawet miliona dolarów, lecz dane zebrane przez Sedera mówiły, że końscy astmatycy nigdy nie robią dobrych czasów na torze, zatrudnił więc asystenta, który siedział na mecie i wyławiał ich spośród startujących.

Z tysiąca koni na aukcji w Ocala wszystkie testy Sedera przeszłoby pozytywnie może około dziesięciu. Jeff w ogóle nie zwraca uwagi na rodowód, bierze pod uwagę wyłącznie wpływ, jaki drzewo genealogiczne może wywrzeć na cenę zwierzęcia. „Rodowód mówi nam, że taki czy inny koń ma znikome szanse na wielki sukces – twierdzi. – Ale jeżeli widzę na własne oczy, że jest świetny, to co mnie może obchodzić, kim byli jego przodkowie?”

Pewnego wieczoru zaprosił mnie do swojego pokoju w hotelu Hilton w Ocala, wyciągnął zdjęcia żony, córki i syna i opowiedział mi o swoim dzieciństwie, rodzinie oraz drodze do zawodu. Był jednym z trzech żydowskich uczniów w liceum w Filadelfii; gdy zdawał maturę, mierzył

cztery stopy i dziesięć cali wzrostu (na studiach dobił do pięciu stóp i dziewięciu cali). Opowiedział mi też o swoim ulubionym koniu, który nosił nazwę Pinky Pizwaanski. Seder kupił go i sam nadał mu to imię na cześć pewnego gejowskiego dżokeja. Uważał, że Pinky – koń – zawsze dawał z siebie wszystko, chociaż nie należał do najlepszych na torze.

W końcu pokazał mi plik komputerowy zawierający wszystkie dane o koniu numer 85, dzięki któremu osiągnął największy sukces w życiu. Czy wyjawiał mi sekret? Być może, ale jak powiedział, nic go to nie obchodziło. Ważniejsze niż ochrona tajemnicy było dla niego to, że jak ostatecznie udowodnił, miał rację; chciał pokazać światu, że warto było przez dwadzieścia lat obcinać nogi padłym zwierzętom, szuflować łajno i własnoręcznie konstruować ultrasonografy.

Oto kilka informacji o koniu numer 85:

Percentyle Numeru 85 (później American Pharoah) jako jednorocznia

	Percentyl
Wzrost	56
Waga	61
Rodowód	70
Lewa komora serca	99,61

Można więc było zobaczyć czarno na białym, dlaczego Seder i jego zespół dostali takiej obsesji na punkcie Numeru 85: ze względu na percentyl lewej komory jego serca!

Poza tym wszystkie pozostałe organy wewnętrzne tego konia, w tym śledziona, również były wyjątkowo duże. Jeff odkrył, że im większa lewa komora serca, tym na ogół dla konia wyścigowego lepiej, chociaż lewa komora takich rozmiarów może także być oznaką choroby, jeżeli inne narządy wewnętrzne są niewspółmiernie małe. American Pharoah

wszystkie najważniejsze organy miał jednak większe od przeciętnych, a lewą komorę olbrzymią. Z analizy danych wynikało, że Numer 85 to koń jeden na sto tysięcy, a może nawet na milion.

Czego więc analitycy mogą się nauczyć dzięki badaniom Sedera?

Po pierwsze i najważniejsze, jeżeli zamierzacie używać nowych danych, by zrewolucjonizować jakąś dziedzinę nauki, najlepiej jest zrobić to na takim polu, gdzie stare metody po prostu się nie sprawdzają. Obsesyjnie przywiązani do rodowodów znawcy, których pokonał Seder, pozostawili mnóstwo miejsca na rozmaite ulepszenia, podobnie jak nastawione wyłącznie na liczenie słów wyszukiwarki, które pokonał Google.

Pewną słabością Google'a, jeśli chodzi o prognozowanie zachorowań na grypę[17], jest to, że można już zupełnie dobrze przewidywać ich liczbę, po prostu stosując dane z ostatniego tygodnia i dokonując sezonowych korekt. Wciąż dyskutuje się o tym, jak wiele wnoszą do tego prostego i potężnego modelu informacje dotyczące wyszukiwań. Moim zdaniem Google ma większe szanse, gdy chodzi o schorzenia, o których pozyskujemy mniej danych, dlatego Google STD[6*] może okazać się w ostatecznym rozrachunku przydatniejszy niż Google Flu.

Drugi wniosek, jaki możemy wyciągnąć z dokonań Sedera, brzmi tak, że nie za bardzo trzeba się martwić o to, co sprawia, że wasza teoria się sprawdza. Jeff nie potrafił mi dokładnie wyjaśnić, dlaczego lewa komora serca ma tak poważny wpływ na wyniki konia wyścigowego, nie był też w stanie powiedzieć, dlaczego taka istotna jest śledziona. Być może pewnego dnia końscy kardiolodzy i hematolodzy rozwikłają te zagadki, na razie jednak nie ma to znaczenia. Seder zajmuje się prognozowaniem, nie wyjaśnianiem. A w prognozowaniu trzeba po prostu umieć przewidzieć, że coś się stanie – nie trzeba wiedzieć dlaczego.

Na przykład Walmart korzysta z danych o sprzedaży we wszystkich swoich sklepach, żeby ustalić, jakie produkty powinny się znajdować na ich półkach. Przed atakiem huraganu Frances, który uderzył na południowo-wschodnie wybrzeże Stanów Zjednoczonych w 2004 roku,

Walmart podejrzewał – całkiem słusznie – że w oczekiwaniu na atak niszczycielskiego wiatru ludzie chyba zmienią trochę zwyczaje konsumenckie, pracownicy firmy przejrzeli zatem dane o zakupach, jakie robili mieszkańcy przed uderzeniami wcześniejszych huraganów. Co kupowano najczęściej? Ciastka Pop-Tarts o smaku truskawkowym[18]. Przed nadejściem wichury ten właśnie produkt sprzedaje się nawet siedem razy lepiej niż zazwyczaj.

Opierając się na takiej analizie, Walmart zamówił całe ciężarówki tych ciastek i przetransportował je drogą międzystanową 95 do swoich sklepów w rejonie, gdzie miał uderzyć huragan. I rzeczywiście sprzedawały się dobrze.

Dlaczego akurat one? Może dlatego, że nie trzeba ich przechowywać w lodówce ani piec. A z jakiego powodu konsumenci preferują smak truskawkowy? Nie mam pojęcia. Wygląda jednak na to, że kiedy przychodzi kataklizm, sięgają najchętniej właśnie po te ciastka. Odtąd przed atakiem huraganu Walmart zawsze zapełnia swoje półki pudełkami truskawkowych pop-tartów. I nie jest ważne, z czego wynika ten związek, liczy się bowiem on sam. Może któregoś dnia uczeni od żywienia czy dietetycy zrozumieją, co łączy silny wiatr i kruche ciastka z dżemem truskawkowym, tymczasem jednak w oczekiwaniu na takie wyjaśnienie Walmart wciąż musi zapełniać półki swoich sklepów pop-tartami, gdy zbliża się sezon sztormów, rice krispies zachować natomiast na bardziej słoneczne dni.

Do podobnych wniosków jasno prowadzi też historia Orleya Ashenfeltera, ekonomisty z Uniwersytetu Harvarda, który jest takim autorytetem wśród winiarzy, jak Seder wśród znawców koni.

Jeszcze nieco ponad dziesięć lat temu Ashenfelter miał zmartwienie. Kupował dużo czerwonego wina z regionu Bordeaux we Francji – niekiedy było pyszne, warte wysokiej ceny, częściej jednak jego jakość przynosiła mu rozczarowanie.

Dlaczego, zastanawiał się Ashenfelter, trzeba płacić tyle samo za mniej smaczne wino?

Któregoś dnia dostał pewną wskazówkę od znajomego dziennikarza i konesera win. Okazało się, że można przewidzieć, czy wino będzie

dobre. Kluczem do jego smaku, powiedział Ashenfelterowi ów znajomy, jest pogoda panująca w okresie dojrzewania winorośli.

Ashenfeltera zaintrygowała ta informacja. Próbował ustalić, czy jest prawdziwa i czy będzie mógł odtąd już zawsze kupować tylko dobre wino. Ściągnął sobie na komputer dane o pogodzie w rejonie Bordeaux z ostatnich trzydziestu lat. Zapisywał także ceny na aukcjach, na które wino trafia zwykle wiele lat od daty pierwotnej sprzedaży, bo dzięki nim można się zorientować, jakiej ostatecznie jest jakości.

Rezultaty tych dociekań były zdumiewające. W ogromnej większości przypadków smak wina zależał właśnie od pogody panującej w okresie dojrzewania.

Właściwie jakość wina da się sprowadzić do jednego prostego równania, które możemy nazwać Pierwszym Prawem Uprawy Winorośli:

Cena = 12,145 + 0,00117 opadów zimowych + 0,0614 średniej temperatury w okresie dojrzewania – 0,00386 opadów w czasie zbiorów.

Ale dlaczego jakość win z regionu Bordeaux zależy od tych czynników? Jak moglibyśmy wyjaśnić Pierwsze Prawo Uprawy Winorośli? Istnieje przynajmniej częściowe wytłumaczenie dla wzoru Ashenfeltera na dobre wino – wysoka temperatura i wilgotność gleby na wczesnym etapie wzrostu są niezbędne, by winogrona odpowiednio dojrzały.

Szczegóły tego predykcyjnego wzoru wykraczają jednak znacznie poza jakąkolwiek teorię i prawdopodobnie nie zostaną nigdy w pełni zrozumiane nawet przez znawców przedmiotu.

Dlaczego jeden centymetr zimowego deszczu zwiększa przeciętnie cenę butelki dojrzałego czerwonego wina o jedną dziesiątą centa? Dlaczego nie o dwie dziesiąte? Albo nie o pięć? Na te pytania nikt nie potrafi na razie odpowiedzieć. Wiadomo tylko, że jeśli w zimie spadnie dodatkowo tysiąc centymetrów deszczu, będziesz musiał zapłacić dodatkowego dolara za butelkę.

Ashenfelter nie wiedział zbyt dobrze, dlaczego jego regresja przynosi takie wyniki, ale kierował się nimi w wyborze win. I jak utrzymuje, „rezultaty są świetne”[19]. Jakość wina, które pija dzisiaj, wyraźnie się podniosła.

Jeśli chcesz przewidzieć przyszłość – wiedzieć, które wino będzie smaczne, które produkty będą się dobrze sprzedawać i które konie będą szybko biegać – nie musisz się za bardzo przejmować tym, dlaczego twoja teoria działa tak, jak działa. Powinieneś tylko wszystko poprawnie obliczyć. I to jest druga lekcja, jakiej udzielił nam Jeff Seder na przykładzie swojej przygody z końmi.

Ostatni wniosek, jaki mi się nasuwa na podstawie zakończonych powodzeniem wysiłków Sedera, by wytypować potencjalnego zwycięzcę Potrójnej Korony, brzmi: trzeba być otwartym i elastycznym, jeżeli chodzi o to, co uważamy za dane. Nie było przecież tak, że zanim pojawił się Seder, eksperci końscy starszej daty w ogóle lekceważyli dane analityczne – przeciwnie, pilnie studiowali rodowody koni wyścigowych i ich czasy na torze. Genialność pomysłu Sedera polegała na tym, żeby szukać tych danych tam, gdzie nikt inny jeszcze ich nie szukał, żeby wziąć pod uwagę niekonwencjonalne źródła informacji. Świeże i oryginalne spojrzenie może przynieść analitykowi wyłącznie korzyści.

SŁOWA JAKO DANE

Pewnego dnia w 2004 roku dwóch młodych ekonomistów, specjalistów od mediów, a w tamtym czasie doktorantów Harvardu, czytało doniesienia prasowe o niedawnej decyzji sądu w Massachusetts legalizującej małżeństwa homoseksualne.

Matt Gentzkow i Jesse Shapiro, bo tak nazywali się ci ekonomiści, zauważyli coś interesującego: że dwie gazety napisały o tym samym wydarzeniu uderzająco innym językiem. „Washington Times”, znany jako dziennik konserwatywny, zatytułował artykuł *Homoseksualiści „biorą ślub” w Massachusetts*, podczas gdy „Washington Post”, mający opinię

gazety liberalnej, donosił, że jest to dzień zwycięstwa dla „par jednopłciowych".

Nie ma się co dziwić, że różne wydawnictwa prasowe mogą mieć różne preferencje i że gazety mogą opisywać to samo wydarzenie z różnych punktów widzenia. Gentzkow i Shapiro zastanawiali się już wcześniej, czy mogliby wykorzystać swoje ekonomiczne wykształcenie do tego, aby zrozumieć światopoglądową stronniczość mediów. Dlaczego niektóre dzienniki i czasopisma prezentują bardziej liberalne, a inne bardziej konserwatywne opinie?

Gentzkow i Shapiro nie wiedzieli jednak, jak podejść do tego pytania. Nie potrafili opracować metody, dzięki której mogliby systematycznie i obiektywnie badać subiektywizm środków masowego przekazu.

W artykule o homoseksualnym ślubie zaciekawiło ich jednak głównie nie to, że dwie gazety przedstawiły tę sprawę inaczej, tylko to, na czym polegała różnica w ich opisie tego wydarzenia – a sprowadzała się ona do wyraźnej odmienności w doborze słów. W 2004 roku słowo „homoseksualiści", którego użył „Washington Times", było już w języku angielskim staroświeckim i nieco lekceważącym określeniem gejów, a wyrażenie „para jednopłciowa", użyte w „Washington Post", podkreślało fakt, że związki gejowskie to po prostu inna forma relacji uczuciowej.

Młodzi naukowcy zastanawiali się, czy język może być kluczem do zrozumienia stronniczości. Czy liberałowie i konserwatyści konsekwentnie używają różnych sformułowań? Czy słownictwo, stosowane w artykułach prasowych, może stanowić zbiór danych? Czego mogą dowieść takie badania amerykańskiej prasy? Skąd możemy wiedzieć, czy dana gazeta jest liberalna, czy konserwatywna? I czy uda nam się zrozumieć dlaczego? W 2004 roku nie były to pytania bezsensowne. Miliardy słów w amerykańskich gazetach zostały już wtedy uwolnione z więzienia druku albo mikrofilmu. Niektóre portale zapisywały teraz pełne teksty wszystkich artykułów publikowanych w niemal każdym tytule prasowym w Stanach Zjednoczonych, Gentzkow i Shapiro mogli je więc przeglądać i szybko zmierzyć stopień, w jakim język wpływał na stronniczość prasy. Mogli też dzięki temu

przyczynić się do lepszego zrozumienia sposobów działania mediów informacyjnych.

Ale zanim powiemy, co zdołali ustalić, zostawmy na moment uczonych oraz ich próbę kwantyfikacji języka prasowego, porozmawiajmy natomiast o tym, w jaki sposób badacze z różnych dyscyplin używają tego nowego dla nich rodzaju danych – słów – po to, aby poszerzyć naszą wiedzę o naturze ludzkiej.

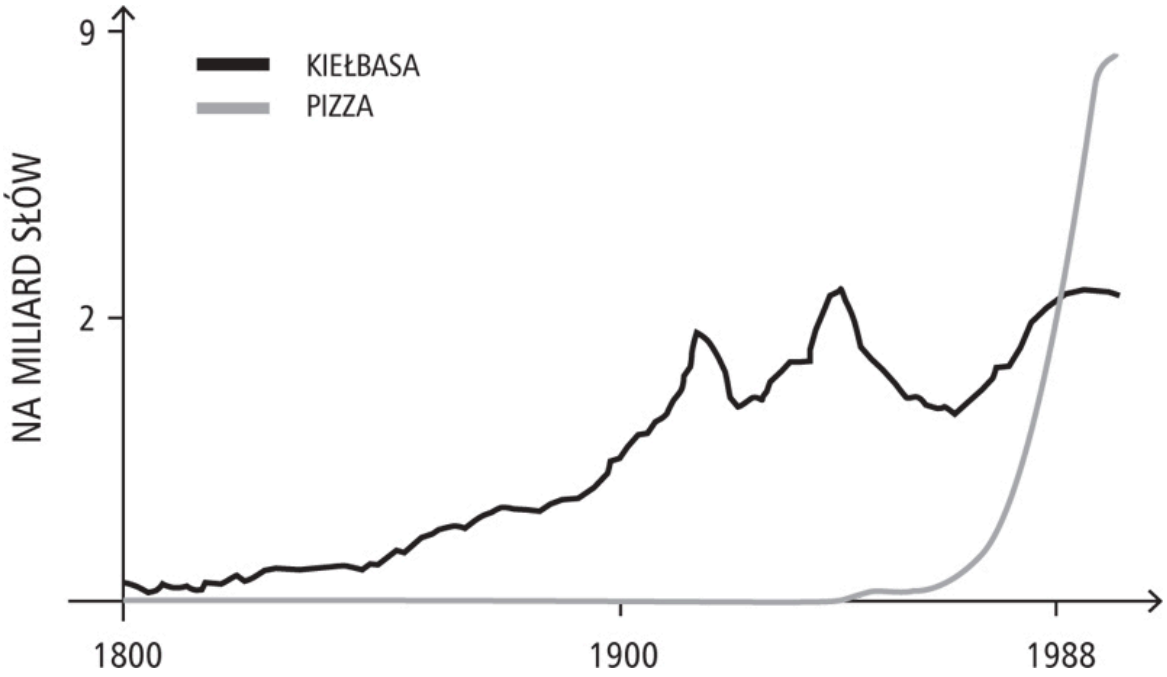
Oczywiście język zawsze interesował przedstawicieli nauk społecznych, badania językoznawcze zwykle wymagały jednak uważnej lektury tekstów, a przekształcanie ich gigantycznych porcji w dane naukowe nie było dawniej możliwe. Dziś dzięki komputerom oraz cyfryzacji wyszukiwanie określonego słownictwa w wielkich bazach tekstowych jest łatwe. W ten sposób język stał się obiektem badań big data. Linki wykorzystywane przez Google'a składają się ze słów, podobnie jak wpisy w tej wyszukiwarce, które analizuję. Niniejsza książka również składa się przede wszystkim ze słów. Język stał się niezwykle istotnym składnikiem rewolucji big data i zasługuje na osobny podrozdział. Wykorzystuje się go w analizach tak często, że powstała nowa poświęcona mu dziedzina badawcza: tekst jako zbiór danych.

Ważnym krokiem na tym polu jest Google Ngrams. Przed kilku laty dwóch młodych biologów, Erez Aiden i Jean-Baptiste Michel, zatrudniło swoich asystentów do tego, żeby liczyli jedno po drugim słowa w starych, zakurzonych tekstach, usiłując odkryć, w jaki sposób rozpowszechnia się znaczenie i sposób zastosowania pewnych wyrazów. Któregoś dnia Aiden i Michel dowiedzieli się o nowym projekcie Google'a, zmierzającym do cyfrowego zapisu znacznej części biblioteki światowej – a językoznawcy niemal od razu pojęli, że dzięki temu będzie im znacznie łatwiej zrozumieć historię języka.

„Zdaliśmy sobie sprawę, że nasze metody są beznadziejnie anachroniczne – powiedział Aiden czasopismu «Discover». – Było jasne, że nie możemy konkurować z gigantem cyfryzacji". On i Michel postanowili więc podjąć współpracę z firmą komputerową i z pomocą inżynierów z Google'a stworzyli program, który przepatruje miliony

zapisanych cyfrowo książek w poszukiwaniu konkretnego słowa albo frazy i informuje badaczy, jak często to słowo czy fraza powtarzały się każdego roku w latach 1800–2010.

Czego możemy się na tej podstawie dowiedzieć? Na przykład tego, jak wolno rosła popularność kiełbasy, a jak szybko i gwałtownie rozpowszechniła się pizza.



Istnieją jednak i o wiele poważniejsze wnioski. Google Ngrams może nam powiedzieć między innymi, jak powstawała tożsamość narodowa. Jeden taki fascynujący przykład pojawia się w książce Aidena i Michela, *Uncharted*.

Ale najpierw szybkie pytanie. Czy waszym zdaniem USA są dzisiaj krajem zjednoczonym czy raczej podzielonym? Jeśli myślicie tak jak większość z nas, powiecie, że ostatnio chyba podzielonym, a to ze względu na wysoki stopień polaryzacji politycznej. Być może nawet uznacie, że tak bardzo podzielonym jak zawsze. Dzisiejsza Ameryka jest dwukolorowa: stany czerwone są republikańskie, a stany niebieskie – demokratyczne. Ale w *Uncharted* Aiden i Michel wskazują na pewien frapujący szczegół, który dowodzi, że dawniej nasz kraj był znacznie bardziej rozdarty. Ten szczegół to język, jakim ludzie mówią o swojej ojczyźnie.

Zwróćcie uwagę na słowa, których użyłem w powyższym akapicie, pytając o podział kraju. Napisałem „The United States *is* divided…" – czyli potraktowałem Stany Zjednoczone jak rzeczownik w liczbie pojedynczej. To w angielszczyźnie naturalna, poprawna gramatycznie i powszechnie akceptowana forma. Jestem pewien, że nawet tego nie zauważyliście.

Ale Amerykanie nie zawsze tak mówili i na początku dziejów swojego kraju używali liczby mnogiej. Na przykład John Adams w orędziu o stanie Unii z roku 1799 mówił o „Stanach Zjednoczonych i *ich* traktatach z Jego Brytyjską Mością" [„the United States in *their* treaties with his Britanic Majesty"]. Gdybym pisał w roku 1800, użyłbym więc formy „The United States *are* divided". Ta drobna różnica w liczbie rzeczownika od dawna fascynowała historyków, wskazuje bowiem na to, że przyszedł taki moment, kiedy Ameryka przestała myśleć o sobie jako o zlepku poszczególnych stanów i zaczęła uważać się za jeden naród.

A kiedy to nastąpiło? Jak wynika z *Uncharted*, historycy nigdy nie mieli co do tego pewności, nie było bowiem sposobu, aby udzielić naukowej odpowiedzi na takie pytanie. Wiele osób podejrzewało jednak od dawna, że do zmiany przyczyniła się wojna secesyjna. James McPherson, były prezes American Historical Association i laureat Nagrody Pulitzera, stwierdził bez ogródek: „Wojna wyznaczyła zmianę liczby rzeczownika «United States» z mnogiej na pojedynczą".

Okazuje się jednak, że McPherson się mylił, a Aiden i Michel mogli to wykazać dzięki Google Ngrams. Mogli na przykład sprawdzić, jak często amerykańskie książki używały nazwy własnej „USA" jako rzeczownika w liczbie mnogiej, a jak często w liczbie pojedynczej w każdym roku istnienia Stanów Zjednoczonych. Otóż zmiana ta następowała bardziej stopniowo i przyspieszyła dopiero już dość długo po zakończeniu wojny secesyjnej.

Jeszcze piętnaście lat po wojnie secesyjnej nazwę „United States" traktowano częściej jako rzeczownik w liczbie mnogiej, co wskazuje, że Stany Zjednoczone były wtedy nadal podzielone językowo. Zwycięstwa militarne zajmują mniej czasu niż zmiany w sposobie myślenia.

* * *

Tyle jeśli chodzi o to, w jaki sposób może dojść do zjednoczenia kraju. A jak łączą się ze sobą ludzie? Słowa mogą nam dopomóc również w odpowiedzi na to pytanie.

Możemy na przykład przewidzieć, czy mężczyzna i kobieta umówią się na drugą randkę, analizując sposób, w jaki rozmawiają ze sobą na pierwszej.

Dowiódł tego interdyscyplinarny zespół uczonych z uniwersytetów Stanforda i Northwestern: Daniel McFarland, Dan Jurafsky i Craig Rawlings. Przebadali oni setki heteroseksualnych „szybkich randkowiczów"[20], próbując ustalić, na jakiej podstawie można by przypuszczać, że na pierwszym spotkaniu przypadną sobie do gustu i będą chcieli zobaczyć się powtórnie.

McFarland, Jurafsky i Rawlings skorzystali najpierw z danych tradycyjnych. Zapytali randkowiczów o wzrost, wagę i zainteresowania, po czym sprawdzili, jak te czynniki korelują z kimś, kto zdradza chociażby cień chęci nawiązania z nimi romantycznego stosunku. Generalizując, kobiety wolą mężczyzn wyższych i takich, którzy podzielają ich zainteresowania, mężczyźni natomiast preferują kobiety szczuplejsze – oraz takie, które podzielają ich zainteresowania. To nic nowego.

Ale uczeni zebrali także nowy rodzaj danych. Poprosili randkowiczów, żeby wzięli ze sobą na spotkanie dyktafony, a nagrania z randek zostały następnie zapisane cyfrowo. Dzięki temu można było potem ponumerować wszystkie słowa, jakich używali na spotkaniu, odnotować śmiechy i ton głosu. Można było też sprawdzić, jak mężczyźni i kobiety sygnalizowali, że są zainteresowani dalszą znajomością, oraz czym zasłużyli sobie na zainteresowanie partnera lub partnerki.

Co powiedziały nam dane językowe? Po pierwsze, w jaki sposób mężczyźni i kobiety ujawniają, że osoba, z którą się umówili, ich pociąga. Czasem jest to zupełnie oczywiste: on śmieje się z jej żartów. Ale czasami te sygnały są mniej ewidentne: na przykład mężczyzna w trakcie rozmowy utrzymuje ton głosu przez cały czas mniej więcej na tym samym poziomie. Istnieją badania, które wskazują na to, że kobiety często postrzegają jednostajny głos jako męski, można zatem przypuszczać, że faceci, być może podświadomie, pragną w ten sposób podkreślić swoją męskość w obecności kobiety, która im się podoba.

Tymczasem, jak się okazało, panie okazują zainteresowanie partnerom, używając różnych tonów głosu, mówiąc ciszej i krócej od nich. Ważna wskazówka to także dobór słownictwa. Kobieta prawdopodobnie nie jest zainteresowana mężczyzną, jeśli używa tak zwanych o(d)gródek (*hedge words*), czyli nieostrych słów i zwrotów w rodzaju „chyba", „jakby" czy „być może".

Panowie – jeżeli kobieta „o(d)gradza się" za pomocą *hedge words* od odpowiedzialności za to, co mówi, i to bez względu na temat rozmowy, jeśli „chyba" smakuje jej drink, jest jej „jakby" chłodno albo „być może"

zje jeszcze jedną przystawkę, to można się założyć, że „chyba”, „jakby” i „być może” nie jesteście w jej typie.

Jeżeli jednak opowiada o sobie, to prawdopodobnie jest zainteresowana mężczyzną. Wygląda na to, że najpiękniejszym słowem, jakie może usłyszeć facet z ust kobiety, z którą chce nawiązać bliższy kontakt, jest „ja” – bo świadczy ono o tym, że kobieta czuje się swobodnie. Jej zainteresowania miałyby dowodzić także takie zwroty fatyczne, jak „Rozumiesz?”, „Wiesz?” czy „Chodzi mi o to, że…”. Dlaczego? Uczeni uważają, że są one prośbą o uwagę ze strony słuchacza. Brzmią przyjaźnie, ciepło i sugerują, że mówca ma ochotę nawiązać z nami kontakt; wiecie, co mam na myśli?

No dobrze, a jak mają się porozumiewać kobiety i mężczyźni, by zainteresować sobą partnerów i partnerki? Dane analityczne świadczą, że jest wiele sposobów, dzięki którym mężczyzna może werbalnie zwiększyć swoje szanse, by przypaść kobiecie do gustu. Panie lubią na przykład takich mężczyzn, którzy dają się im prowadzić. Może nie ma w tym nic zaskakującego, ale kobieta częściej przyzna, że facet jest w jej typie, jeśli będzie się śmiał z jej dowcipów i podtrzymywał rozmowę na tematy, które proponuje ona, zamiast ciągle zmieniać je na takie, o których sam chce rozmawiać[7*]. Lubią też facetów, którzy okazują im wsparcie i współczucie. Jeśli mężczyzna mówi: „To niesamowite!” albo „To naprawdę cool!”, istnieje większe prawdopodobieństwo, że kobieta się nim zainteresuje. Podobnie jeśli będzie używał takich zwrotów jak „To ciężka sprawa” albo „Więc pewnie jest ci smutno”.

Przynoszę jednak i trochę kiepskich wiadomości dla pań, bo dane zdają się potwierdzać pewną niesmaczną prawdę o mężczyznach. Rozmowa odgrywa jedynie niewielką rolę w ich stosunku do kobiet. Najważniejszym elementem predykcyjnym, wskazującym na to, czy osobnik mojej płci zwróci uwagę na kobietę, jest jej aparycja. Płeć piękna może nieznacznie zwiększyć swoje szanse tylko jednym słowem, o którym już wspominaliśmy: „ja”. Jest bowiem po prostu tak, że faceci częściej lgną do pań, które mówią o sobie. I odwrotnie – po randce, na której rozmawiały z mężczyzną na swój temat, kobiety częściej przyznają, że chciałyby ją powtórzyć. Z tego wynika, że to ważny sygnał, jeśli w czasie pierwszego spotkania konwersacja w znacznej mierze

krąży wokół kobiety, która ujawnia w ten sposób, że czuje się na luzie, i docenia to, że facet nie dominuje nad nią werbalnie. Z kolei mężczyźnie podoba się, że kobieta się przed nim otwiera, ich druga randka jest zatem prawdopodobna.

Na koniec dodajmy, że w randkowej transkrypcji można też znaleźć przynajmniej jedną zapowiedź problemów – a jest nią znak zapytania. Im więcej pytań pada podczas pierwszego rendez-vous, tym mniej prawdopodobne, że zarówno kobieta, jak i mężczyzna wyrażą zainteresowanie drugim. Ten sygnał wydaje się kontrintuicyjny, bo można by przypuszczać, że pytania oznaczają ciekawość. Owszem, ale nie na pierwszej randce – ponieważ wtedy w większości znamionują znudzenie. „Jak spędzasz wolny czas?" „Czy masz rodzeństwo?" Pytamy o takie rzeczy, kiedy rozmowa zamiera. Udana pierwsza randka powinna się kończyć tylko w jeden sposób: „Umówisz się ze mną jeszcze?". Jeżeli jest to jedyne pytanie, jakie padło w czasie naszego pierwszego spotkania, odpowiedź najprawdopodobniej będzie brzmiała: „Tak".

Poza tym mężczyźni i kobiety mówią inaczej nie tylko wtedy, kiedy usiłują się nawzajem poderwać, ale w ogóle, na co dzień.

Zespół psychologów przeanalizował słownictwo z setek tysięcy postów na Facebooku. Zbadano, jak często mężczyźni i kobiety używają poszczególnych jednostek leksykalnych, dzięki czemu można było ogłosić, które wyrazy są najbardziej „zmaskulinizowane", a z których najchętniej korzystają kobiety.

Niestety wiele z tych preferencji było oczywistych. Na przykład kobiety mówią o „zakupach" i swoich „włosach" o wiele częściej niż mężczyźni, a ci z kolei częściej opowiadają o „futbolu" i „xboksach" niż panie, więc pewnie nie trzeba było całego zespołu psychologów do analizy big data, by wam to uświadomić.

Niektóre z ustaleń uczonych okazały się jednak ciekawsze. Na przykład kobiety znacznie częściej używają słowa „jutro" – może dlatego, że mężczyźni nie są zbyt dobrzy w wybieganiu myślami w przyszłość. Jedną z najbardziej kobiecych cech językowych jest też wydłużanie dźwięku „o" w słówku „so" [„więc", „a zatem"], stąd wśród

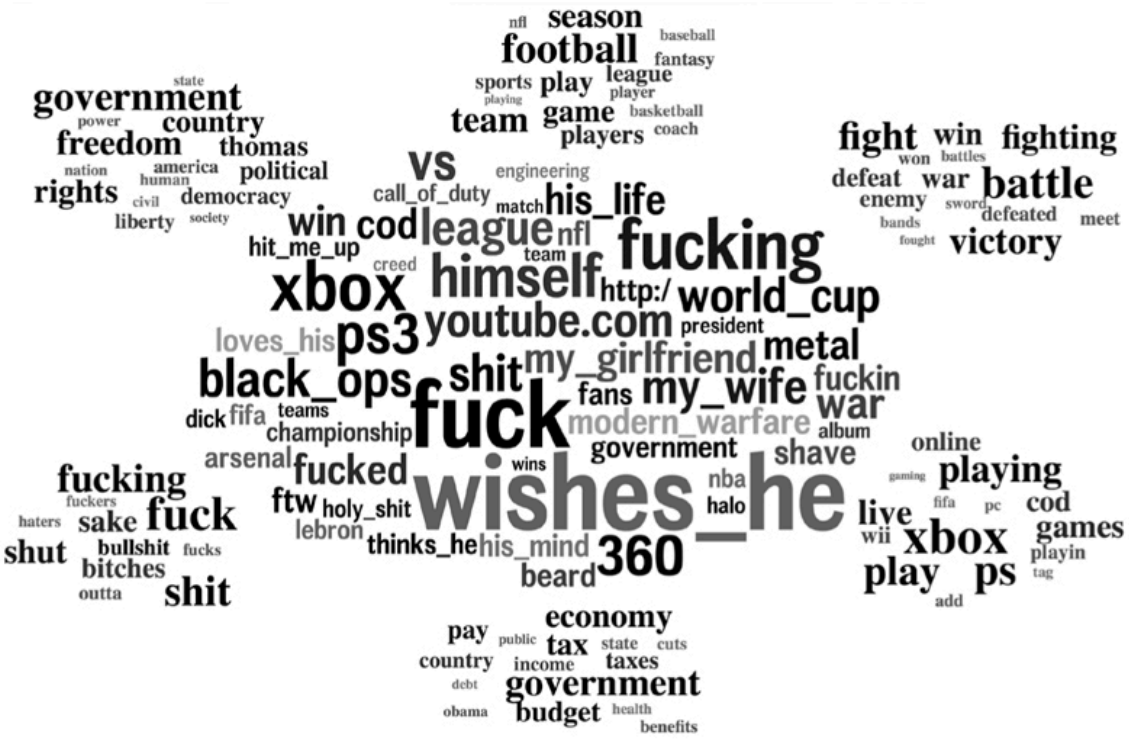
wyrazów używanych nieproporcjonalnie często przez panie znajdują się „soo", „sooo", „soooo", „sooooo" i „soooooo".

Może dlatego, że wychowywałem się wśród kobiet, które nie boją się od czasu do czasu rzucić mięsem, myślałem zawsze, że obie płci przeklinają mniej więcej równie często, ale tak nie jest. Do słów używanych znacznie częściej przez mężczyzn niż kobiety zaliczają się „fuck" [„kurwa mać", „ja pierdolę", „jebać"], „shit" [„gówno", „cholera"], „fucks" [„o kurwa", „kurde", „zjeby"], „bullshit" [„gówno prawda"], „fucking" [„jebany", „pierdolony"] i „fuckers" [„chuje", „skurwysyny"].

Oto „chmury" słów zawierające wyrazy używane najczęściej przez mężczyzn i kobiety. Im większy rozmiar pisma, tym chętniej z danego wyrazu korzystają przedstawiciele jednej lub drugiej płci (zob. s. 109).

Podoba mi się w tym badaniu to, że nowe dane umożliwiają nam poznanie pewnych istniejących już od dawna norm, o których na ogół nie mieliśmy pojęcia. Mężczyźni i kobiety zawsze mówili niejako różnymi językami, ale przez setki tysięcy lat materiały źródłowe na ten temat znikały z chwilą rozpłynięcia się fal dźwiękowych w przestrzeni. Dzisiaj przechowuje się je w komputerach i można je analizować cyfrowo.

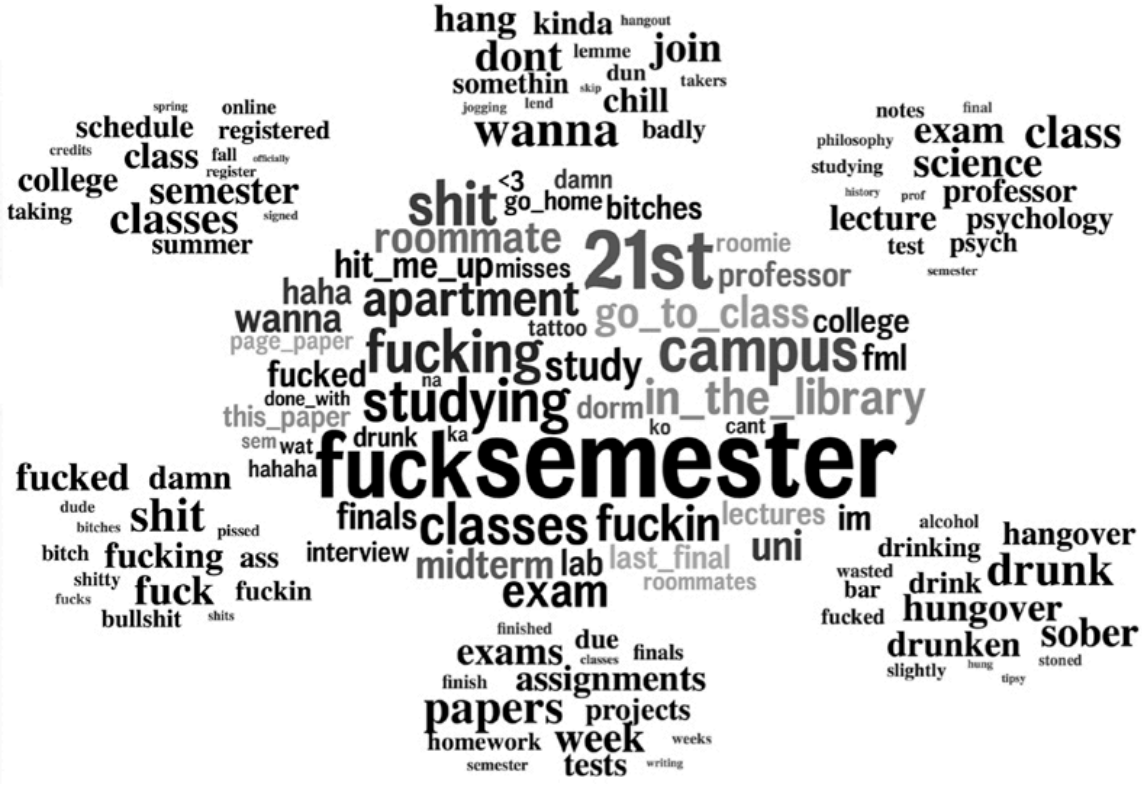
MĘŻCZYŹNI


nfl
season
football
baseball
fantasy
sports
play
league
player
playing
team
game
basketball
players
coach
state
government
power
country
freedom
thomas
nation
america
human
political
rights
civil
democracy
liberty
society
vs
engineering
call_of_duty
match
his_life
win
cod
league
nfl
fucking
hit_me_up
team
creed
xbox
himself
http:/
world_cup
youtube.com
president
loves_his
ps3
my_girlfriend
metal
black_ops
shit
fans
my_wife
fuckin
war
dick
fifa
teams
championship
fuck
modern_warfare
album
government
shave
online
arsenal
fucked
wins
wishes
_he
nba
playing
gaming
fucking
fuckers
ftw
holy_shit
halo
fifa
pc
cod
haters
sake
fuck
lebron
live
wii
xbox
games
playin
shut
bullshit
fucks
thinks_he
his_mind
360
bitches
beard
play
ps
tag
outta
shit
add
economy
pay
public
tax
state
cuts
country
income
taxes
government
debt
obama
budget
health
benefits
fight
win
fighting
won
battles
defeat
war
battle
enemy
sword
defeated
meet
bands
fought
victory

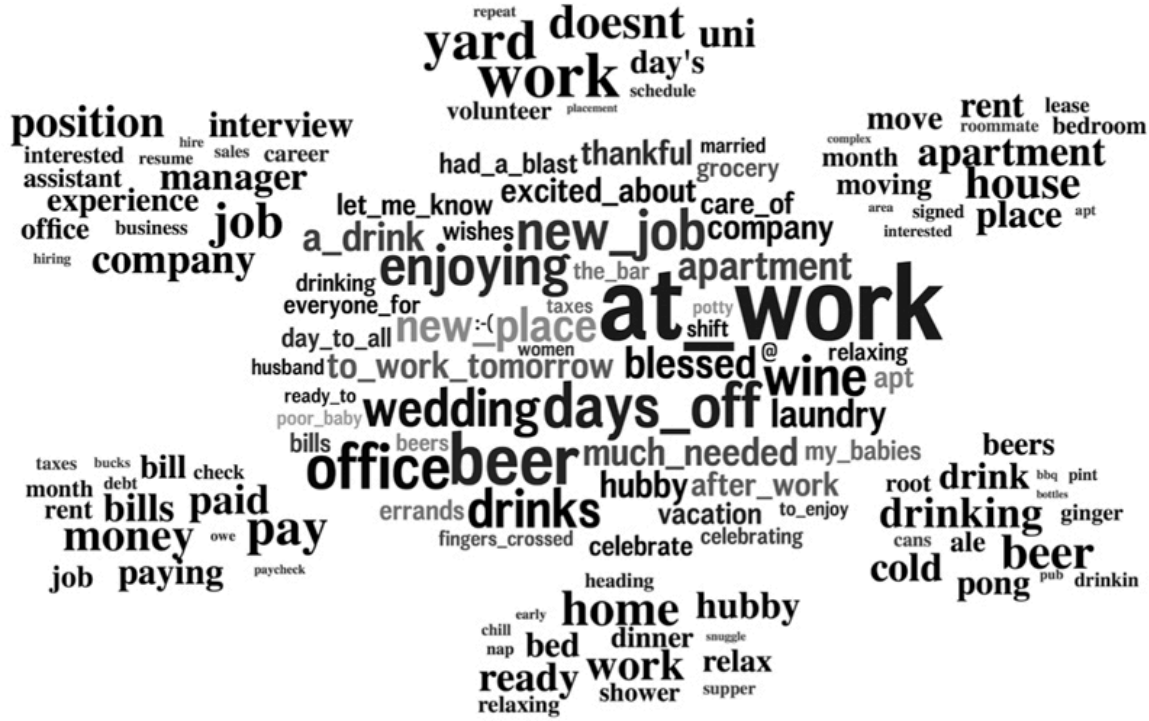
KOBIETY

Biorąc jednak pod uwagę moją płeć, powinienem chyba powiedzieć teraz tak: „Słowa, kurwa ich mać, kiedyś po prostu znikały. A teraz możemy zrobić sobie przerwę od oglądania futbolu i grania na xboksie i poczytać sobie o tym całym syfie. Oczywizda, jeżeli kogoś to w ogóle, kurwa, coś obchodzi”.

OSOBY W WIEKU 19–22 LAT



OSOBY W WIEKU 23–29 LAT



OSOBY W WIEKU 30–65 LAT

Nie tylko mężczyźni i kobiety posługują się różnymi językami. Ludzie mówią też inaczej w zależności od wieku. Być może dzięki temu dowiemy się więcej o tym, jak przebiega proces starzenia się. Poniżej podaję wyniki badań tych samych naukowców oraz grupy wyrazów najczęściej używanych przez osoby w różnym wieku na Facebooku. Tytułuję te grafiki: „Picie”, „Praca”, „Modlitwa”. Bo nastolatki piją. Dwudziestolatkowie pracują. Trzydziestolatkowie i ludzie starsi się modlą.

***

Nowy potężny instrument do badania tekstów nosi czasem nazwę „analizy emocjonalnej”. Naukowcy potrafią dzisiaj ustalić, w jakim stopniu dany fragment tekstu jest radosny lub smutny.

Jak się to robi? Zespoły uczonych proszą sporą liczbę osób o to, by określiły znaczenie dziesiątek tysięcy słów w języku angielskim jako pozytywne albo negatywne. Najbardziej pozytywne słowa, jak mówią wyniki badań, to między innymi „szczęśliwy”, „miłość” i „niesamowity”,

a do najbardziej negatywnych zaliczają się „smutny", „śmierć" i „depresja". W ten sposób naukowcy skonstruowali indeks nastrojów wielkiego zbioru wyrazów.

Stosując taki indeks, można mierzyć średnie uczuciowe natężenie słów w tekście. Jeśli ktoś pisze: „Jestem szczęśliwy, zakochany i czuję się niesamowicie", to analiza uczuciowa oznaczy to wyznanie jako tekst skrajnie radosny. Jeżeli ktoś pisze natomiast: „Jest mi smutno, kiedy myślę o śmierci i depresji na świecie", to analiza uczuciowa zakwalifikuje ten tekst jako skrajnie smutny. Inne wypowiedzi pisemne mieściłyby się gdzieś pomiędzy tymi dwoma biegunami.

Czego możemy się dowiedzieć, wyznaczając nastrój tekstu? Analitycy danych badający Facebooka wskazali jedną ekscytującą możliwość. Otóż da się ocenić codzienny Stan Szczęśliwości Narodowej. Jeżeli status większości wpisów będzie pozytywny, można przypuszczać, że kraj jest tego dnia szczęśliwy; jeżeli negatywny – że jest smutny.

Naukowcy badający wpisy na Facebooku ustalili między innymi, że jeden z najradośniejszych dni w roku to Boże Narodzenie. Przyznaję, że byłem sceptycznie nastawiony do tych analiz i w ogóle podchodzę nieufnie do powyższego projektu. Ogólnie rzecz biorąc, uważam, że wielu ludzi w głębi duszy smuci się w święta Bożego Narodzenia, ponieważ są samotni albo skłóceni ze swoimi bliskimi. Rzecz biorąc zaś jeszcze ogólniej, z powodów, które omówię bliżej w następnym rozdziale, mam ograniczone zaufanie do wpisów na Facebooku, a to głównie dlatego, że kiedy opowiadamy w sieci o swoim życiu, to zwykle kłamiemy.

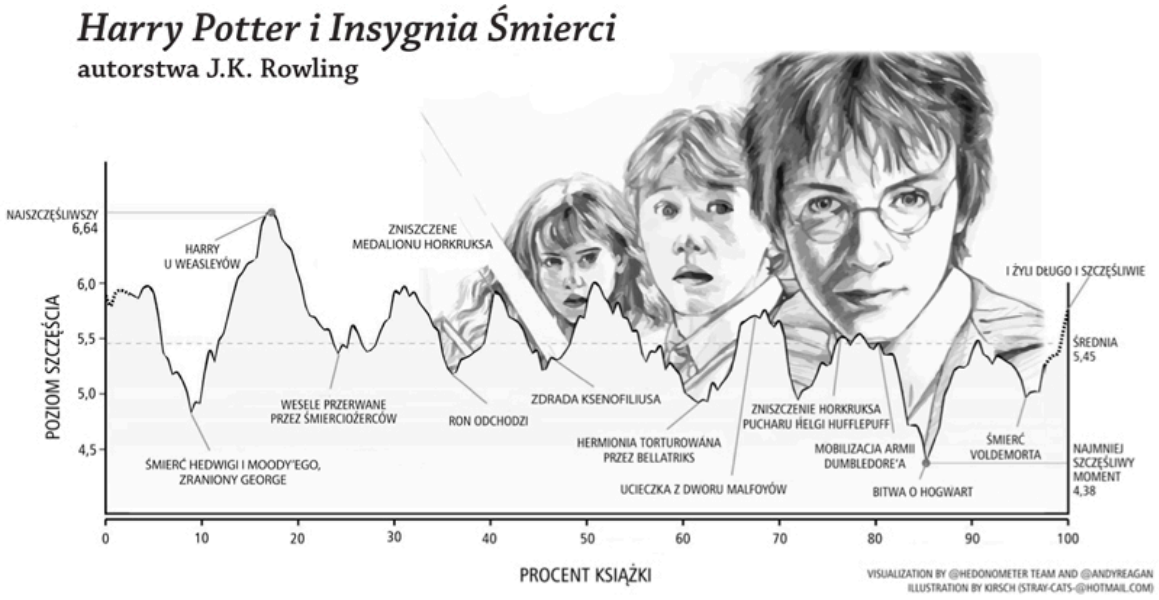
Jeżeli jesteś samotny i przygnębiony w święta, to czy naprawdę masz ochotę zawracać głowę wszystkim swoim znajomym, pisząc, jaki się czujesz nieszczęśliwy? Podejrzewam, że wiele osób spędza Boże Narodzenie w ponurym nastroju, chociaż na Facebooku wyrażają wdzięczność wobec losu za swoje „cudowne, niesamowite, zdumiewające i radosne życie" – a wtedy takich internautów oznacza się jako autorów, którzy znacznie podnoszą aktualny Stan Szczęśliwości Narodowej. I jeżeli naprawdę zamierzamy mierzyć coś takiego, powinniśmy korzystać z większej liczby źródeł niż wpisy na Facebooku.

Z tym zastrzeżeniem można chyba jednak powiedzieć, że twierdzenie o radosnym charakterze świąt Bożego Narodzenia jest generalnie prawdziwe. Wyszukiwania w Google'u na temat depresji i sondaże Gallupa także wskazują, że jest to jeden z najradośniejszych okresów w roku. A wbrew miejskim legendom liczba samobójstw w okolicach świąt maleje. Owszem, znajdziemy w Boże Narodzenie ludzi smutnych i samotnych, przeważnie jednak jesteśmy wtedy weseli.

W dzisiejszych czasach jeśli ludzie już w ogóle coś czytają, to ich lekturą są przeważnie wpisy na portalach społecznościowych. Ale kiedyś, nie tak dawno temu, istoty ludzkie czytywały różne opowieści, czasami w formie książek. Analiza uczuć może nas wiele nauczyć również pod tym względem.

Zespół naukowców z Andym Reaganem na czele, obecnie pracownikiem University of California w Berkeley School of Information, zapisywał cyfrowe wersje tysięcy książek i scenariuszy filmowych[22], aby następnie określić, na ile radosne lub smutne są ich punkty zwrotne.

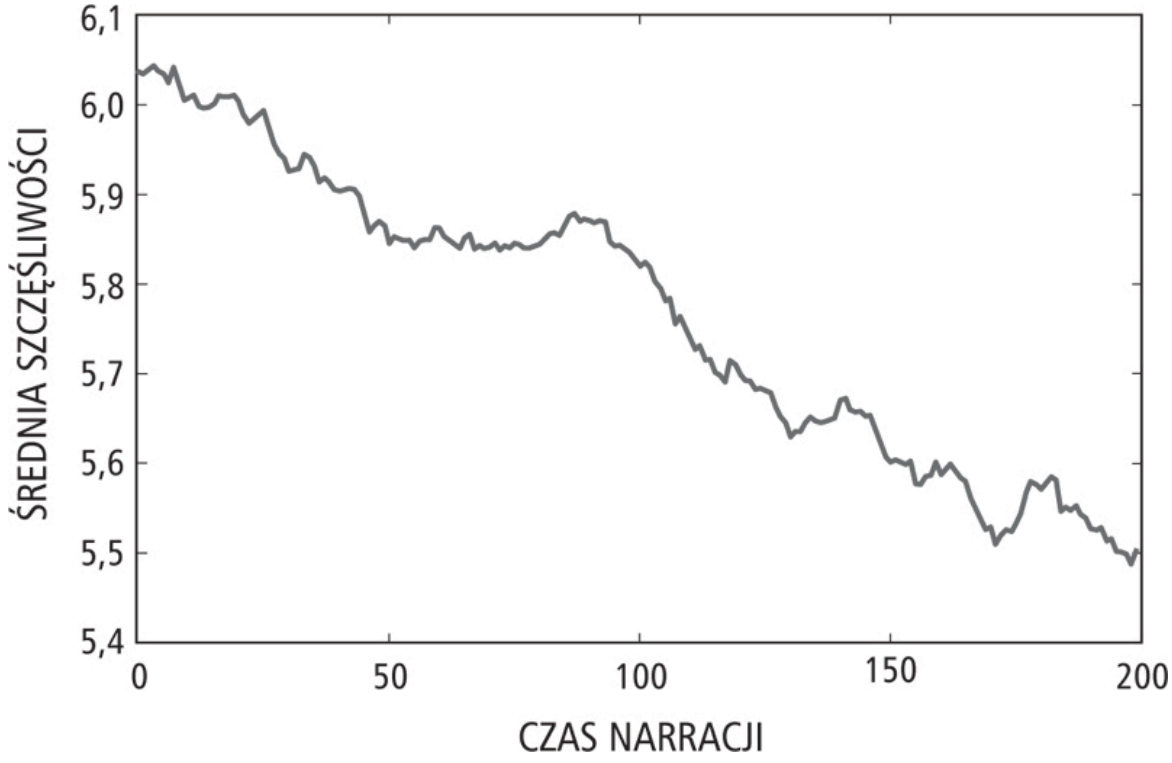
Weźmy na przykład powieść *Harry Potter i Insygnia Śmierci*. Na wykresie poniżej uczeni pokazują, jak zmienia się nastrój fabuły, wyszczególniając opisowo jej najważniejsze zdarzenia dramaturgiczne.



Zwróćcie uwagę na to, że liczne nastrojowe wzloty i upadki, które wskazuje analiza uczuciowa, są ściśle powiązane z kluczowymi wydarzeniami fabularnymi.

Większość fabuł odznacza się prostą budową. Spójrzmy na tragedię Shakespeare'a *Życie i śmierć króla Jana*. W tej sztuce nic nie układa się pomyślnie dla bohatera. Żąda się od niego, aby się zrzekł tronu. Papież nakłada na niego ekskomunikę za nieposłuszeństwo. Wybucha wojna. Umiera siostrzeniec Jana, być może popełniając samobójstwo. Giną też inne osoby z jego otoczenia. Na końcu król Jan zostaje otruty przez gderliwego mnicha.

Poniższy wykres przedstawia analizę emocjonalną tego dramatu w miarę rozwoju akcji.

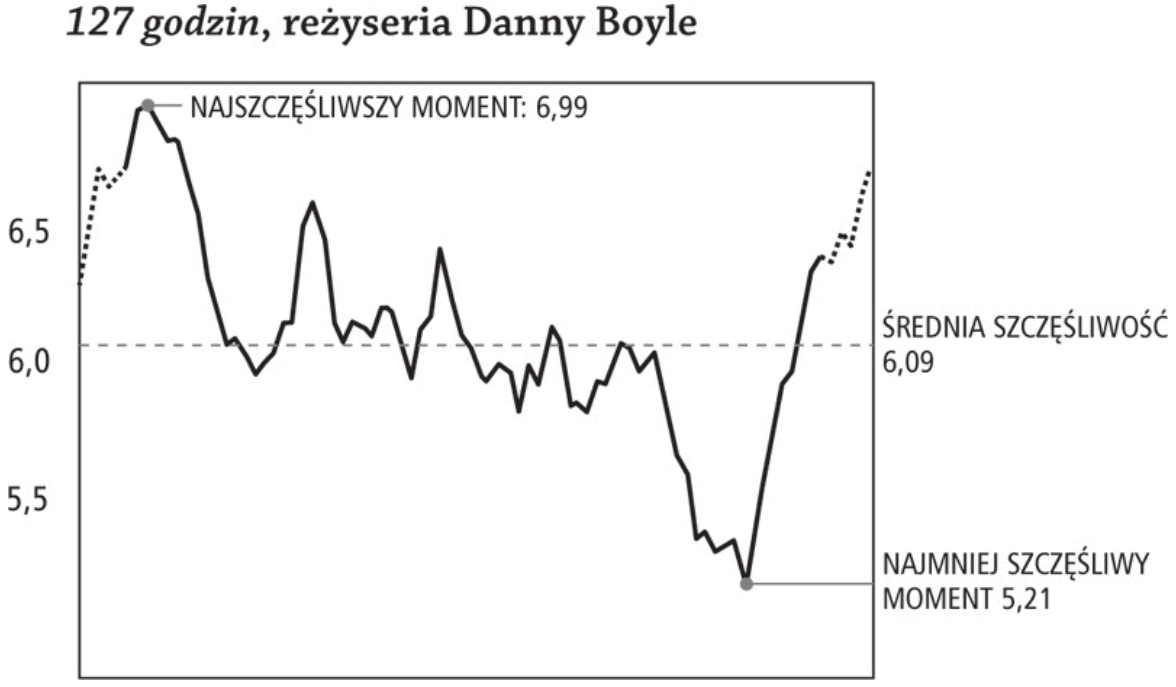


Okazuje się więc, że na podstawie samej analizy leksykalnej komputer potrafił określić, że sytuacja bohatera ze złej staje się gorsza, a potem jeszcze gorsza.

Albo przyjrzyjmy się filmowi *127 godzin*. Jego krótkie streszczenie wygląda następująco:

Alpinista jedzie na wspinaczkę do Utah, do Canyonlands National Park. Spotyka innych wspinaczy, ale w końcu ich drogi się rozchodzą. W pewnym momencie traci równowagę i potrąca głaz, który zaklinowuje w szczelinie skalnej jego dłoń i nadgarstek. Próbuje wydostać się z pułapki na różne sposoby, lecz bez powodzenia. Popada w depresję. W końcu odcina sobie dłoń i w ten sposób wydostaje się z matni. Żeni się, zakłada rodzinę i wspina się dalej, teraz jednak przed wyjściem w góry zawsze zostawia wiadomość, dokąd się udaje.

Poniżej zamieszczam wykres analizy emocjonalnej filmu, dokonanej przez zespół uczonych Reagana.



Czego możemy się nauczyć na podstawie analizy nastrojów tysięcy fabuł książkowych i filmowych?[23].

Informatycy dowiedli, że ogromna większość opowieści da się sprowadzić do sześciu stosunkowo prostych struktur. Zgodnie z tabelą ułożoną przez zespół Reagana są to:

Od pucybuta do milionera (wzlot)

Od milionera do pucybuta (upadek)

Człowiek w dołku (upadek, wzlot)

Ikar (wzlot, upadek)

Kopciuszek (wzlot, upadek, wzlot)

Edyp (upadek, wzlot, upadek)

Ten prosty schemat może nie uwzględniać drobnych zawiłości fabuły; na przykład film *127 godzin* jest sklasyfikowany jako opowieść typu „człowiek w dołku", chociaż znajdziemy tu chwile, kiedy nastrój na pewien czas się podnosi. Lecz podstawowa budowa dramaturgiczna większości fabuł podpada pod jedną z wymienionych sześciu kategorii. *Harry Potter i Insygnia Śmierci* to wyjątek.

Pozostaje jednak wiele dodatkowych pytań, na które moglibyśmy spróbować odpowiedzieć. Na przykład w jaki sposób struktura opowieści zmieniała się w czasie? Czy z biegiem lat ich fabuły się komplikowały? Czy kultury różnią się między sobą, jeśli chodzi o typ ich ulubionych opowieści? Jakie fabuły ludzie lubią najbardziej? Czy mężczyźni wolą inne fabuły niż kobiety? Jakie opowieści lubią mieszkańcy jednego kraju, a jakie drugiego?

Tak czy owak, tekst jako zbiór danych może nam dostarczyć bezprecedensowych informacji o tym, jakich historii oczekuje publiczność, i mogą to być fabuły inne niż te, których chcą widzowie i czytelnicy zdaniem zarówno pisarzy i scenarzystów, jak i szefów wytwórni i wydawnictw. Mamy już pewne wskazówki co do preferencji odbiorców.

Spójrzmy na badania przeprowadzone przez dwoje profesorów Wharton School, Jonaha Bergera i Katherine L. Milkman, a dotyczące opowieści „zaraźliwych", którymi ludzie dzielą się z sobą w sieci. Sprawdzili między innymi, czy na listę najczęściej wysyłanych e-mailem artykułów z „New York Timesa" trafiają raczej teksty o wydźwięku pozytywnym, czy jednak negatywnym. Zbadali pod tym kątem wszystkie artykuły z tej gazety publikowane w okresie trzech miesięcy i zgodnie z metodologią analizy emocjonalnej wyznaczali ich nastrój. Przykłady pozytywnych tekstów to *Wide-Eyed New Arrivals Falling in Love with the City* [*Nowi przybysze robią wielkie oczy i zakochują się w mieście*] i *Tony Award for Philanthropy* [*Nagroda Tony za filantropię*]. Natomiast

artykuły takie jak *Web Rumors Tied to Korean Actress' Suicide* [*Plotki w sieci miały przyczynić się do samobójstwa koreańskiej aktorki*] czy *Germany: Baby Polar Bear's Feeder Dies* [*Niemcy: Zmarł opiekun niedźwiadka polarnego*] zostały uznane, co nie powinno dziwić, za teksty o wydźwięku negatywnym.

Uczeni wiedzieli też, gdzie artykuły zostały wydrukowane. Czy na stronie głównej? W prawym górnym rogu? W lewym górnym? No i wiedzieli, kiedy tekst się ukazał. We wtorek późnym wieczorem? Czy w poniedziałek rano?

Zestawiali zatem ze sobą dwa artykuły – pozytywny z negatywnym – które ukazały się w internetowym wydaniu „Timesa" w podobnym miejscu i czasie, po czym sprawdzali, który z nich czytelnicy częściej rozsyłali pocztą elektroniczną do swoich znajomych[24].

No to jak, chcecie wiedzieć, czy chętniej dzielimy się pozytywnymi, czy negatywnymi tekstami?

Pozytywnymi. Autorzy badań konkludują: „Im bardziej pozytywny emocjonalnie jest tekst, tym większe prawdopodobieństwo, że stanie się popularny".

Proszę zwrócić uwagę, że ten wniosek zdaje się kolidować z rozpowszechnioną opinią dziennikarską, głoszącą, że ludzie chętniej czytają opowieści katastroficzne i pełne przemocy. Jest zapewne prawdą, że media informacyjne dostarczają odbiorcom mnóstwa mrocznych historii. I jest trochę prawdy w krążącym po newsroomach powiedzeniu: „Jak coś we krwi tonie, to na pierwszą stronę". Badania uczonych z Wharton School sugerują jednak, że ludzie mogą chcieć weselszych historii, co skłania mnie do ukucia nowego powiedzonka: „Jak coś się uśmiecha, to trafia do poczty elektronicznej" – chociaż ta maksyma wcale się nie rymuje.

Tyle jeśli chodzi o teksty smutne i radosne. A w jaki sposób sprawdzić, czy dane słownictwo jest liberalne albo konserwatywne? I co nam to mówi o współczesnych mediach informacyjnych? To trochę bardziej skomplikowana sprawa, cofająca nas do badań Gentzkowa i Shapiro. Jak pamiętacie, to ci ekonomiści, którzy zauważyli, że to samo

gejowskie małżeństwo zostało opisane w różny sposób w dwóch różnych gazetach, a potem próbowali odpowiedzieć na pytanie, czy badając leksykę wypowiedzi, uda się ujawnić jej stronniczość.

Nasi dwaj młodzi i ambitni naukowcy zbadali najpierw transkrypcje Congressional Record [Archiwów Kongresu]. Zostały one już wcześniej zdigitalizowane, Gentzkow i Shapiro mogli więc zapisać na twardych dyskach każde słowo, którego używali demokratyczni i republikańscy kongresmeni w 2005 roku, a następnie stwierdzić, czy istnieją takie sformułowania, z których demokraci korzystają chętniej niż republikanie i odwrotnie.

Okazało się, że tak. Oto kilka przykładów:

Sformułowania używane znacznie częściej przez demokratów	**Sformułowania używane znacznie częściej przez republikanów**
podatek spadkowy	podatek od śmierci
prywatyzacja ubezpieczeń społecznych	reforma ubezpieczeń społecznych
Rosa Parks	Saddam Husajn
prawa pracownicze	prawo własności prywatnej
ludzie ubodzy	wydatki rządowe

Co tłumaczy różnicę w sposobach wypowiedzi przedstawicieli obu partii?

Zdarza się, że demokraci i republikanie stosują różne określenia do opisania tego samego pomysłu. W 2005 roku republikanie próbowali obniżyć federalny podatek spadkowy i chętnie mówili o nim jako o „podatku od śmierci" (co sugeruje poniekąd, że podatek nakłada się na zmarłego). Demokraci przedstawiali go jako podatek spadkowy (co sugeruje poniekąd podatek od wzbogacenia). Republikanie usiłowali też przenieść ubezpieczenia społeczne na indywidualne konta emerytalne i nazywali tę propozycję reformą, chociaż dla demokratów była to raczej niebezpiecznie brzmiąca „prywatyzacja".

Czasami różnice w języku wypowiedzi sprowadzają się do kwestii nacisku na jej niektóre elementy. Należy chyba przypuszczać, że

zarówno republikanie, jak i demokraci mają ogromny szacunek dla Rosy Parks, bohaterki ruchu na rzecz praw obywatelskich, lecz demokraci wspominali ją częściej. Podobnie ma się sprawa z Saddamem Husajnem, byłym przywódcą Iraku, bo najprawdopodobniej i republikanie, i demokraci sądzą, że był on zbrodniczym dyktatorem – ale to republikanie wymieniali go częściej, próbując usprawiedliwić wojnę w Iraku. Z kolei „prawa pracownicze" i troska o „ludzi ubogich" są problemami zasadniczymi dla Partii Demokratycznej, tak jak „prawo własności prywatnej" i cięcia „wydatków rządowych" dla republikanów.

A omawiane tu różnice leksykalne w wypowiedziach polityków obu partii są znaczne. W 2005 roku republikańscy kongresmeni użyli określenia „podatek od śmierci" trzysta sześćdziesiąt pięć razy, natomiast „podatek spadkowy" tylko czterdzieści sześć. W wypadku demokratów proporcje są odwrotne. Określenia „podatek od śmierci" użyli trzydzieści pięć razy, a „podatek spadkowy" sto dziewięćdziesiąt pięć.

Naukowcy zdali sobie sprawę, że skoro dobór słów może nam podpowiedzieć, który kongresmen jest demokratą, a który republikaninem, to można na tej podstawie również się zorientować, czy dana gazeta ma bardziej lewicowy, czy prawicowy charakter, jeżeli bowiem kongresmeni republikańscy używają sformułowania „podatek od śmierci" najprawdopodobniej po to, aby przekonać ludzi, że należy mu się przeciwstawić, to podobnie mogą się zachowywać konserwatywne gazety. I tak relatywnie liberalny dziennik „Washington Post" używał określenia „podatek spadkowy" trzynaście i siedem dziesiątych razy częściej niż „podatek od śmierci", konserwatywny „Washington Times" wykorzystywał natomiast oba te określenia mniej więcej w tej samej liczbie.

Dzięki cudom Internetu Gentzkow i Shapiro mogli przeanalizować słownictwo znacznej liczby amerykańskich tytułów prasowych. Wykorzystywali dwa portale, newslibrary.com i proquest.com, które razem skatalogowały cyfrowo czterysta trzydzieści trzy dzienniki. Aby ustalić, jakie są sympatie polityczne tych gazet, uczeni obliczyli, jak często występuje w nich wybrany tysiąc nacechowanych politycznie

słów i sformułowań. Okazało się, że najbardziej liberalną gazetą w Stanach jest „Philadelphia Daily News", a najbardziej konserwatywną „Billings Gazette", wychodząca w stanie Montana.

Kiedy mamy już pierwsze wyniki badań profili mediów, reprezentatywne dla tak dużej liczby podmiotów, możemy chyba pokusić się o próbę odpowiedzi na zapewne najważniejsze pytanie: dlaczego niektóre gazety zdradzają sympatie lewicowe, inne zaś prawicowe?[25]

Nasi ekonomiści szybko zauważyli jeden kluczowy czynnik: dominujące w danym regionie poglądy polityczne. Jeżeli są one na ogół liberalne, jak w okolicach Filadelfii i Detroit, to główne tamtejsze gazety skłaniają się ku liberalizmowi, jeżeli natomiast mówimy o bardziej konserwatywnych miastach, jak Billings albo Amarillo w Teksasie, to należy się spodziewać, że okoliczne dzienniki będą konserwatywne. Wynika z tego wyraźnie, że prasa zazwyczaj oferuje swoim czytelnikom takie treści, jakich oczekują.

Można by pomyśleć, że na polityczne inklinacje takiego czy innego tytułu prasowego wpływa jego właściciel, okazuje się jednak, że jego znaczenie jest w tym względzie mniejsze, niż sądzimy. Proszę zwrócić uwagę, co się dzieje, jeśli jedna osoba czy też firma zostaje właścicielem kilku dzienników, wychodzących w różnych miejscach. Spójrzmy na przykład na New York Times Company. Spółka ta wydaje liberalny, jak dowodzą Gentzkow i Shapiro, dziennik „New York Times" w Nowym Jorku, gdzie blisko siedemdziesiąt procent mieszkańców ma poglądy demokratyczne – ale w czasie, kiedy prowadzono te badania, jej własnością był też skłaniający się, zdaniem obu uczonych, ku konserwatyzmowi „Spartanburg Herald-Journal" w Spartanburgu w Karolinie Południowej, gdzie blisko siedemdziesiąt procent mieszkańców to zwolennicy republikanów. Są oczywiście wyjątki: „New York Post", tytuł chyba przez wszystkich uważany za konserwatywny, należy do News Corporation Ruperta Murdocha[26]. Na ogół jednak badania wykazują, że o politycznych sympatiach gazet bardziej decyduje rynek niż właściciel.

Badania te mogą poważnie wpłynąć na to, w jaki sposób myślimy o środkach przekazu. Wiele osób, zwłaszcza o poglądach

marksistowskich, uważa, że amerykańskie dziennikarstwo znajduje się w rękach bogaczy lub wielkich korporacji, których celem jest propaganda wśród mas, rzekomo prowadzona po to, by przekonywać ludzi do poglądów politycznych właścicieli gazet. Artykuł Gentzkowa i Shapiro dowodzi jednak, że nie jest to ich główna motywacja. Właściciele amerykańskich tytułów prasowych dają czytelnikom przede wszystkim to, czego chcą, a postępują tak dlatego, by dzięki temu zwiększyć swoje dochody.

Och, i jeszcze jedno pytanie – ważne, kontrowersyjne, a może nawet bardziej prowokujące. Czy przeciętnie rzecz biorąc, amerykańskie gazety skłaniają się na lewo, czy na prawo? Czy nasze media informacyjne są bardziej liberalne czy konserwatywne?

Zdaniem Gentzkowa i Shapiro amerykańska prasa skłania się raczej na lewo. Nasze gazety używają języka częściej przypominającego język kongresmenów demokratycznych.

„A więc tak! – zakrzyknie na to zapewne odbiorca prawicowy. – A nie mówiłem?" Wielu konserwatystów od dawna podejrzewa, że prasa generalnie jest stronnicza i manipuluje czytelnikami w taki sposób, by popierali poglądy lewej strony sceny politycznej.

Tak nie jest, mówią autorzy, a liberalizm gazet stanowi odpowiedź na sympatie czytelników, którzy w nieznacznej większości skłaniają się na lewo (Gentzkow i Shapiro dysponowali odpowiednimi danymi na ten temat).

Nie istnieje zatem żaden wielki spisek. Jest tylko kapitalizm.

Wyniki wspomnianych tu badań pokazują, że media informacyjne funkcjonują podobnie jak każde inne przedsiębiorstwo na naszym globie. Supermarkety sprawdzają, jakich klienci chcą lodów, a następnie zapełniają nimi swoje półki, gazety natomiast dowiadują się, jakich czytelnicy oczekują od nich poglądów, i zapełniają nimi swoje łamy. „To jest po prostu biznes", powiedział mi Shapiro[27]. Oto czego można się dowiedzieć, jeśli skwantyfikować tak skomplikowane zjawiska jak wiadomości, analizy i opinie i rozłożyć je na czynniki pierwsze, czyli słowa.

OBRAZY JAKO DANE

Zazwyczaj kiedy uczeni akademiccy albo przedsiębiorcy potrzebowali danych z jakiejś dziedziny, przeprowadzali ankiety i sondaże. Informacje docierały do nich starannie sformatowane, czerpano je ze statystyk albo zaznaczanych „ptaszkiem" kratek na kwestionariuszach. Dziś już się tego nie robi. Minęły czasy uporządkowanych, czystych i prostych danych, opartych na wynikach badania opinii publicznej. Dziś głównym źródłem informacji stają się brudne ślady, które pozostawiamy za sobą w drodze przez życie.

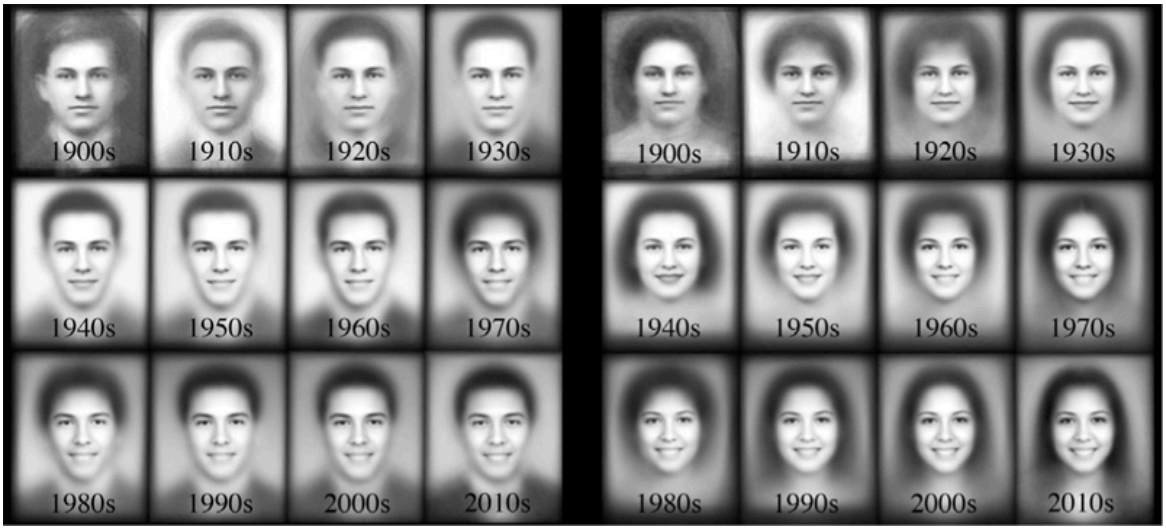
Jak już widzieliśmy, słowa to dane. Danymi są też linki, podobnie jak literówki, banany z naszych snów, ton ludzkiego głosu, rzężenie konia, bicie jego serca, wielkość śledziony. Moim zdaniem zbiór danych przynoszący najbardziej odkrywcze informacje stanowią wpisy w wyszukiwarkach internetowych.

Okazuje się jednak, że danymi są również fotografie.

Słowa, zamknięte kiedyś na zakurzonych półkach w książkach i periodykach, zostały dzisiaj zdigitalizowane, tak jak i zdjęcia, uwolnione z albumów i tekturowych pudełek. Je także przetworzono na bajty i wypuszczono w chmurę. Teksty mogą nam udzielić lekcji historii – na przykład pokazując, jak mówiliśmy kiedyś, a jak mówimy dzisiaj – potrafią to jednak robić również fotografie, chociażby dowodząc, że inaczej pozowaliśmy do nich dawniej, a inaczej pozujemy dzisiaj.

Przyjrzyjmy się pomysłowym badaniom czteroosobowego zespołu informatyków z uniwersytetów Browna i UoC Berkeley, którzy wykorzystali nowe możliwości ery cyfrowej – okazało się bowiem, że wiele liceów w Stanach Zjednoczonych zeskanowało swoje stare fotograficzne albumy szkolne i udostępniło je użytkownikom w sieci. Badacze znaleźli w Internecie dziewięćset czterdzieści dziewięć takich albumów z lat 1905–2013[28], a w nich dziesiątki tysięcy portretów uczniów z klas maturalnych. Za pomocą odpowiedniego oprogramowania uczonym udało się następnie złożyć ze wszystkich zdjęć z danej dekady twarz „przeciętnego" maturzysty. Oznacza to, że wyznaczyli „przeciętne" umiejscowienie i kształt ich nosów, oczu i ust,

jak również „przeciętną” fryzurę. Poniżej widzimy owe „przeciętne” twarze z ostatnich około stu dziesięciu lat; męskie z lewej, kobiece z prawej strony.



Zauważyliście coś ciekawego? Amerykanie – a zwłaszcza Amerykanki – zaczęli się uśmiechać. Na początku XX wieku pozowali z niemal kamiennymi obliczami, a pod koniec niemal promienieją.

Skąd ta zmiana? Czyżby mieszkańcy USA stali się szczęśliwsi?

Nie. Odpowiedzi na to pytanie pomogli informatykom udzielić uczeni z innych dyscyplin – a jest ona, przynajmniej dla mnie, fascynująca. Otóż kiedy wynaleziono fotografię, ludzie traktowali ją jak malarstwo, nie mieli bowiem nic innego, z czym mogliby ją porównać. Dlatego osoby portretowane na zdjęciach upodobniały się do osób portretowanych na obrazach[29]. Ludzie pozujący do portretów nie mogli się ciągle uśmiechać przez długie godziny, jakie zajmowało malowanie, przybierali więc poważne miny – a osoby pozujące do zdjęć ich naśladowały.

Dlaczego jednak zaczęliśmy w końcu zmieniać wyraz naszej fizjonomii? Oczywiście zadziałały tu biznes, zysk i marketing. W połowie XX wieku Kodak, firma z branży filmowej i fotograficznej, zaczęła popadać w kłopoty, gdyż jej klienci robili niewiele zdjęć, wymyśliła więc sposób, by skłonić ich do częstszego fotografowania. Otóż reklamy Kodaka zaczęły kojarzyć zdjęcia z radością i szczęściem, a firma zmierzała do tego, aby ludzie nauczyli się fotografować, kiedy

chcą pokazać innym, jak się dobrze bawią. Wszystkie te uśmiechnięte zdjęcia z albumów szkolnych to nic innego jak wynik skutecznej kampanii reklamowej Kodaka (podobnie jest z większością fotografii, które widzimy dzisiaj na Facebooku albo Instagramie).

Lecz zdjęcia jako dane mogą nam powiedzieć znacznie więcej niż to, kiedy maturzyści zaczęli się uśmiechać, pozując do fotografii. Zapewne zabrzmi to zaskakująco, ale dzięki zdjęciom da się ocenić również stan gospodarki przynajmniej niektórych krajów.

Weźmy taki prowokacyjnie zatytułowany artykuł naukowy: *Jak mierzyć wzrost gospodarczy z orbity okołoziemskiej*. Możecie iść o zakład, że jeśli ktoś taki jak ja natknie się gdzieś na tekst o tego rodzaju tytule, to na pewno go przeczyta. Autorzy wyżej wspomnianego opracowania – J. Vernon Henderson, Adam Storeygard i David N. Weil – zaczynają od stwierdzenia, że w wielu krajach rozwijających się dostępne środki pomiaru produktu krajowego brutto (PKB) są niewystarczające – a to dlatego, że znaczna część aktywności gospodarczej w tych państwach odbywa się w tak zwanej szarej strefie, a agencje rządowe, które powinny badać stan gospodarki, dysponują mocno ograniczonymi środkami finansowymi na swoją działalność.

Na czym polegał niekonwencjonalny pomysł autorów? Na tym, aby PKB w krajach rozwijających się badać na podstawie ilości emitowanego w nich nocą światła[30]. Dane na temat jasności pochodziły ze zdjęć wykonanych przez amerykańskiego satelitę wojskowego, który okrąża Ziemię czternaście razy na dobę.

Dlaczego nocne światło może być dobrym środkiem pomiaru PKB? Bo mieszkańców najuboższych części świata nie stać na płacenie wysokich rachunków za prąd i kiedy sytuacja gospodarcza jest zła, odbiorcy indywidualni, a nawet całe miasteczka, drastycznie obniżają pobór energii elektrycznej po zmroku.

W Indonezji w czasie azjatyckiego kryzysu finansowego w 1998 roku zużycie światła nocą dramatycznie spadło, natomiast w Korei Południowej w latach 1992–2008 zwiększyło się o siedemdziesiąt dwa procent, bo kraj ten przeżywał wtedy niezwykły wzrost gospodarczy. W tym samym okresie w Korei Północnej pobór energii elektrycznej nocą spadł, ponieważ tam sytuacja ekonomiczna stała się fatalna.

W 1998 roku w południowej części Madagaskaru odkryto bogate złoża rubinów i szafirów. Miasteczko Ilakaka, dawniej niewiele więcej niż miejsce odpoczynku dla kierowców ciężarówek, zmieniło się w duży ośrodek handlowy. Przed 1998 rokiem w Ilakace właściwie nie było widać żadnych świateł nocą, ale w następnych pięciu latach dochodziło tam do nocnej eksplozji jasności.

Autorzy przyznają, że ich dane dotyczące nocnego oświetlenia są jako narzędzie pomiaru wyników gospodarczych dalekie od ideału. Z całą pewnością nie można precyzyjnie określić stanu gospodarki, opierając się wyłącznie na tym, ile światła wykrywa satelita po zmroku, a autorzy w ogóle nie zalecają stosowania tej metody w odniesieniu do krajów rozwiniętych, takich jak Stany Zjednoczone, gdzie mamy znacznie dokładniejsze dane ekonomiczne z innych źródeł. Uczciwie mówiąc, ich zdaniem nawet w krajach biedniejszych badanie rzędu jasności światła po zachodzie słońca daje mniej więcej takie same rezultaty jak metody urzędowe. Lecz połączenie niepełnych danych państwowych z niepełnymi danymi wynikającymi z nocnej obserwacji pozwala na lepszą ocenę stanu gospodarki niż każde z tych źródeł z osobna. A zatem tak, korzystając ze zdjęć zrobionych z przestrzeni kosmicznej, możemy lepiej ocenić stan gospodarki krajów rozwijających się.

Joseph Reisinger, obdarzony łagodnym głosem doktor informatyki, podziela zastrzeżenia autorów artykułu co do wartości zbiorów danych na temat potencjału ekonomicznego krajów rozwijających się. Jak pisze, w kwietniu 2014 roku na przykład Nigeria skorygowała swoją wcześniejszą prognozę PKB, uwzględniając w niej nowe, pomijane dotąd sektory produkcyjne. Prognozowane PKB stało się wyższe o dziewięćdziesiąt procent[31].

„Mają największą gospodarkę w Afryce – powiedział Reisinger, stopniowo podnosząc głos. – A my nie wiemy o nich nawet najbardziej podstawowych rzeczy”[32].

Chciał znaleźć jakiś sposób, który pozwoliłby mu dokładniej oceniać wyniki ekonomiczne poszczególnych państw. Pomysł, na który wpadł, stanowi świetny przykład rewizji koncepcji danych analitycznych, jednocześnie dowodząc jej przydatności.

Reisinger założył firmę – Premise – zatrudniającą w krajach rozwijających się uzbrojonych w smartfony pracowników. Ich zadanie? Fotografować wszystko, co może mieć znaczenie gospodarcze.

Smartfonowcy fotografują zatem stacje benzynowe lub kosze z owocami w supermarketach. Robią zdjęcia stale w tych samych miejscach i wysyłają do Premise, gdzie inna grupa pracowników Reisingera – informatycy – przekształca je w dane. Analitycy firmy potrafią odczytać i zinterpretować wszystko, od długości kolejek na stacjach benzynowych przez liczbę jabłek dostępnych w supermarkecie po stopień ich dojrzałości oraz cenę. Na podstawie takich zdjęć Premise może już pokusić się o opinię na temat wyników gospodarczych obserwowanego kraju i wysokości inflacji. W krajach rozwijających się głównym sygnałem problemów są długie kolejki po benzynę, podobnie jak niedojrzałe jabłka w sklepach – albo ich brak (jabłek, nie sklepów). Takie terenowe zdjęcia Premise zrobione w Chinach umożliwiły w 2011 roku wykrycie inflacji, spowodowanej tam głównie cenami żywności, a potem (w roku 2012) deflacji, na długo zanim opublikowano oficjalne dane.

Premise sprzedaje pozyskane tą metodą informacje bankom albo funduszom inwestycyjnym, a także współpracuje z Bankiem Światowym.

Jak wiele dobrych pomysłów, Premise okazał się kurą znoszącą złote jajka. Bank Światowy interesował się niedawno wielkością czarnego rynku papierosów na Filipinach. Chciał zwłaszcza poznać skutki niedawnych działań rządu, w tym nalotów policyjnych, mających na celu likwidację fabryk produkujących papierosy bez akcyzy. Jakie sprytne rozwiązanie zastosowała firma? Jej pracownicy fotografowali paczki papierosów, które widzieli u ludzi na ulicy, by można było sprawdzić, ile z nich jest opatrzonych znaczkami akcyzy, jakich wymagają legalne wyroby tytoniowe. Okazało się, że ta część czarnego rynku, w 2015 roku jeszcze duża, w 2016 znacznie się zmniejszyła. Działania rządu odniosły skutek – ale żeby dostrzec coś tak zazwyczaj starannie ukrywanego, jak czarnorynkowe papierosy, trzeba było zgromadzić nowe dane.

Jak widzimy, w epoce cyfrowej musieliśmy poważnie zrewidować swoje pojęcie o tym, czym są zbiory danych, i okazało się, że w nowych informacjach jest wiele do odkrycia. Powody politycznej stronniczości prasy, udana pierwsza randka i stan gospodarek krajów rozwijających się to ledwie początek.

Nie jest też przypadkiem, że na nowych danych można zbić wielkie pieniądze, takie jak dziesiątki miliardów panów Brina i Page'a. Josephowi Reisingerowi też nieźle poszło. Obserwatorzy szacują, że Premise przynosi dziś rocznie dziesiątki milionów dolarów dochodu, a inwestorzy włożyli ostatnio w firmę pięćdziesiąt milionów[33]. Oznacza to, że przynajmniej niektórzy z nich uważają Premise za jedną z najwięcej wartych agencji fotograficznych na świecie, grającą w tej samej lidze co „Playboy".

A zatem zarówno dla naukowców, jak i przedsiębiorców, wszystkie dostępne obecnie nowe rodzaje danych mają ogromną wartość, także dzięki temu, że do sfery danych zakwalifikowaliśmy nowy typ informacji. W dzisiejszych czasach analitykowi nie wolno ograniczać się do wąskiego czy tradycyjnego poglądu na to, czym są zbiory danych. Wartościowe informacje to współcześnie zdjęcia kolejek w supermarketach, zawartość półek sklepowych i dojrzałość jabłek. Danymi analitycznymi są także fotografie z orbity okołoziemskiej albo kształt ust ludzi portretowanych na zdjęciach. Zbiorem danych może być w naszej dobie dosłownie wszystko.

A mając wielki zbiór do dyspozycji, możemy wreszcie przejrzeć ludzkie kłamstwa.

ROZDZIAŁ CZWARTY
CYFROWE SERUM PRAWDY

Wszyscy kłamią.

Kłamią, mówiąc, ile drinków wypili w drodze do domu. Kłamią, mówiąc, jak często chodzą na siłownię, ile zapłacili za nowe buty, czy przeczytali taką lub inną książkę. Zwalniają się z pracy pod pretekstem choroby, choć wcale nie są chorzy. Mówią, że pozostaną z tobą w kontakcie, chociaż nigdy więcej się do ciebie nie odezwą. Mówią, że to nie dlatego, że cię nie lubią, chociaż to nieprawda. Mówią, że cię kochają, ale nie kochają. Mówią, że są szczęśliwi, kiedy mają doła. Mówią, że podobają im się kobiety, mimo że tak naprawdę wolą mężczyzn.

Okłamujemy przyjaciół. Szefów. Dzieci. Rodziców. Lekarzy. Mężów. Żony. Samych siebie.

I na pewno kłamiemy jak diabli w sondażach i ankietach.

Oto mój krótki sondaż dla was:

Czy kiedykolwiek ściągałaś/ściągałeś na egzaminie?
Czy kiedykolwiek chciałaś/chciałeś kogoś zabić?

Mieliście chęć skłamać w odpowiedzi? Wiele osób ukrywa w sondażach i ankietach swoje wstydliwe zachowania i myśli. Sondaże zwykle są anonimowe, chcemy w nich jednak sprawiać dobre wrażenie. Taką postawę określa się mianem „pragnienia atrakcyjności społecznej" [*social desirability bias*].

Dowodów na to, w jaki sposób sondaże mogą paść ofiarą naszych kłamstw, dostarczył pewien ważny naukowo artykuł już w 1950 roku[1]. Uczeni zebrali wtedy pochodzące z oficjalnych źródeł dane

o mieszkańcach Denver; chcieli się dowiedzieć między innymi, ilu z nich głosowało w wyborach, wspierało pewną miejscową imprezę charytatywną i posiadało kartę biblioteczną. Następnie zadawali pytania mieszkańcom, by sprawdzić, czy uzyskane od nich informacje pokrywają się z oficjalnymi. Wyniki okazały się, jak na tamte czasy, szokujące. Odpowiedzi mieszkańców zasadniczo różniły się od danych, jakie zgromadzili naukowcy. Mimo że ankiety były anonimowe, znaczna liczba pytanych podawała nieprawdę, kiedy chodziło o rejestrację w komisji wyborczej, głosowanie oraz wsparcie dla organizacji dobroczynnej.

	Sondaż	**Dane oficjalne**
Zarejestrowani w komisjach wyborczych	83%	69%
Głos w ostatnich wyborach prezydenckich	73%	61%
Głos w ostatnich wyborach na burmistrza	63%	36%
Karta biblioteczna	20%	13%
Wsparcie dla ostatniej imprezy dobroczynnej Community Chest	67%	33%

Czy w ciągu ostatnich sześćdziesięciu pięciu lat zaszła w tej materii jakaś zmiana? W epoce Internetu brak karty bibliotecznej zapewne przestał być powodem do wstydu, ale chociaż zmieniło się nasze pojęcie tego, co jest wstydliwe albo pożądane, nadal mamy skłonności do oszukiwania ankieterów.

W niedawnym sondażu zapytano absolwentów Uniwersytetu Maryland o przebieg ich studiów, a odpowiedzi porównano z danymi oficjalnymi[2]. Ankietowani notorycznie udzielali fałszywych odpowiedzi po to, żeby polepszyć swój wizerunek publiczny. Mniej niż dwa procent z nich odpowiedziało, że uzyskali dyplom z wynikiem gorszym niż średnia 2,5 (chociaż taki wynik zanotowało jedenaście procent absolwentów). A czterdzieści cztery procent napisało, że

w ostatnim roku wpłaciło dotację na uniwersytet (w rzeczywistości uczyniło to zaledwie dwadzieścia osiem procent badanych).

Na pewno jest możliwe, że kłamstwo odegrało ważną rolę w porażce tych agencji badania opinii publicznej, którym w 2016 roku nie udało się przewidzieć zwycięstwa Donalda Trumpa[3]. Sondaże zaniżały wyniki poparcia dla niego przeciętnie o mniej więcej dwa procent. Być może niektórzy pytani wstydzili się, że zamierzają go poprzeć, inni zaś twierdzili, że są niezdecydowani, chociaż od początku chcieli na niego głosować.

Dlaczego w anonimowych ankietach podajemy informacje niezgodne z prawdą? Spytałem o to Rogera Tourangeau, emerytowanego profesora University of Michigan, zapewne największego eksperta od ludzkiego pragnienia społecznej atrakcyjności. Wyjaśnił mi, że ważną część problemu stanowi nasza słabość do „niewinnych kłamstw". „Na co dzień mniej więcej jedna trzecia tego, co mówimy, to kłamstwa – twierdzi Tourangeau. – I przenosimy te swoje przyzwyczajenia na sondaże"[4].

Mamy też dziwny zwyczaj okłamywania samych siebie. „Jest w ludziach jakaś niechęć do przyznania się przed sobą, że byli, powiedzmy, słabymi studentami", mówi profesor.

Okłamywanie samych siebie wyjaśnia być może, dlaczego tylu z nas przedstawia się jako osoby lepsze niż przeciętne[5]. Czy to poważny problem? Przeszło czterdzieści procent inżynierów z pewnej firmy stwierdziło, że zalicza się do pięciu procent najlepszych specjalistów w swojej dziedzinie. Ponad dziewięćdziesiąt procent profesorów uniwersyteckich utrzymuje, że wartość ich dorobku jest ponadprzeciętna. Jedna czwarta maturzystów uważa, że zalicza się do jednego procenta uczniów, którzy najlepiej dogadują się z innymi ludźmi. Jeżeli oszukujemy nawet samych siebie, to chyba nie możemy być szczerzy w sondażach.

Kolejnym czynnikiem wpływającym na to, że kłamiemy w badaniach opinii publicznej i ankietach, jest nasze silne pragnienie, by wywrzeć dobre wrażenie na nieznajomym, który przeprowadza z nami wywiad, naturalnie jeśli ankieter robi to osobiście. Tourangeau ujmuje to w ten

sposób: „Wchodzi ktoś, kto przypomina ci twoją ulubioną ciotkę… Czy chciałbyś jej powiedzieć, że w zeszłym miesiącu paliłeś marihuanę?[8*] Czy chciałbyś się przyznać, że nie przekazałeś pieniędzy na swoją kochaną, poczciwą Alma Mater?".

Z tego powodu wykazujemy się tym większą szczerością, w im bardziej bezosobowych warunkach się znajdujemy. Dlatego jeśli chcemy uzyskać nieskłamane odpowiedzi, sondaże internetowe będą lepsze niż telefoniczne, a telefoniczne lepsze niż przeprowadzane twarzą w twarz. Ludzie są bardziej szczerzy w samotności niż wtedy, gdy razem z nimi w pokoju znajduje się inna osoba.

W kwestiach wrażliwych jednak żadna metoda sondażu nie zapewni nam w pełni prawdziwych odpowiedzi. Tourangeau użył tu słowa, którym często posługują się ekonomiści: „bodziec". Ludziom brakuje odpowiedniego bodźca, by odpowiadać w sondażach zgodnie z prawdą.

A zatem jak możemy się dowiedzieć, co nasi bliźni naprawdę robią i myślą?

W niektórych wypadkach wystarczą oficjalne źródła danych. Nawet jeśli ludzie kłamią na temat swoich datków na cele dobroczynne, możemy poznać prawdziwą wysokość ich dotacji, zwracając się do organizacji charytatywnych. Ale kiedy próbujemy dowiedzieć się czegoś o ludzkich zachowaniach, których nie odnotowują urzędowe dokumenty, albo co ludzie myślą – poznać ich prawdziwe poglądy, uczucia i pragnienia – nie mamy dostępu do innego źródła informacji niż te odpowiedzi, jakich zapytani zechcą udzielić ankieterom. A raczej nie mieliśmy, aż do tej pory.

Kłania się tu druga moc big data: są takie miejsca w sieci, gdzie ludzie odważają się przyznać do czegoś, do czego nie przyznaliby się nigdzie indziej. Takie miejsca działają jak cyfrowe serum prawdy. Pomyślcie o wyszukiwaniach w Google'u i przypomnijcie sobie, w jakich okolicznościach ludzie stają się bardziej szczerzy. W sieci? Tak. W samotności? Tak. Wtedy, kiedy ankieter nie przeprowadza sondażu osobiście? Tak.

Ale Google ma jeszcze jedną zaletę, która sprawia, że mówimy prawdę: bodźce, motywację. Jeżeli lubisz rasistowskie dowcipy, to nie

masz chyba żadnego dobrego powodu, by podać tę politycznie niepoprawną informację w sondażu – masz jednak motywację, żeby szukać w sieci najlepszych nowych rasistowskich kawałów. Jeżeli wydaje ci się, że cierpisz na depresję, to nie masz motywacji, żeby przyznać się do tego w ankiecie – chociaż masz pobudki, aby zapytać Google'a o symptomy i ewentualne środki lecznicze.

Nawet jeżeli okłamujesz sam siebie, Google może znać prawdę. Na kilka dni przed wyborami ty i niektórzy twoi sąsiedzi możecie być przekonani, że pojedziecie do swojej komisji wyborczej i weźmiecie udział w głosowaniu. Ale jeśli ani ty, ani oni nie wiecie, jak i gdzie głosować, i nie szukaliście tych informacji w Internecie, to analitycy danych, w tym i ja, mogą dojść do wniosku, że frekwencja w waszym okręgu wyborczym okaże się jednak niska. Może też nie przyznałeś się nawet sam przed sobą, że cierpisz na depresję, mimo że szukasz w sieci informacji na temat napadów płaczu i trudności ze wstawaniem rano z łóżka – lecz twoje powiązane z depresją wpisy internetowe zostałyby zarejestrowane pośród innych w okolicy twojego miejsca zamieszkania i można by je przeanalizować, jak to już pokazałem w pierwszej części książki.

Zastanówcie się nad swoimi doświadczeniami z użytkowaniem Google'a. Domyślam się, że od czasu do czasu wpisywaliście w wyszukiwarkę różne rzeczy, ujawniające jakieś zachowania albo myśli, do których w eleganckim towarzystwie wolelibyście się raczej nie przyznawać. Mam przytłaczające dowody na to, że znaczna większość Amerykanów powierza Google'owi wiele bardzo osobistych wyznań. Znacznie częściej szukamy na przykład hasła „porno" niż „prognoza pogody"[7]. Nawiasem mówiąc, trudno pogodzić te dane z informacjami uzyskanymi na podstawie sondaży, bo tylko mniej więcej dwadzieścia pięć procent mężczyzn i osiem procent kobiet przyznaje się do oglądania w sieci treści pornograficznych[8].

Być może zauważyliście także, w jaki sposób wyszukiwarka próbuje automatycznie uzupełnić wpisywane w nią pytania i hasła – jej sugestie opierają się na najczęściej powtarzających się wpisach internetowych, dlatego takie autouzupełnianie pozwala nam się domyślić, czego ludzie szukają. Autouzupełnianie może jednak być także trochę mylące.

Google nie podpowie wam na przykład pewnych słów, które uznaje za nieprzyzwoite, jak „kutas", „jebać" czy „porno"[9], a to oznacza, że na podstawie autouzupełniania możemy sądzić, iż ludzie miewają kosmate myśli rzadziej niż w rzeczywistości. Mimo to wyszukiwarka często podsuwa nam wrażliwe informacje.

Jeżeli wpis zaczniemy od „Dlaczego...", to okaże się, że dwa najczęściej występujące dzisiaj w sieci autouzupełnienia, czyli w rezultacie najczęściej zadawane pytania, brzmią: „...niebo jest niebieskie?" i „...niektóre lata są przestępne?". Trzecie to: „Dlaczego moja kupa ma zielony kolor?". Autouzupełnianie w Google'u bywa też jednak niepokojące. Jeśli wpiszemy na przykład „Czy normalne jest chcieć...", pierwszą automatyczną sugestią okaże się „zabić". A jeżeli wpiszemy „Czy normalne jest chcieć zabić...", to najczęstszą podpowiedzią będzie „swoją rodzinę".

Chcecie więcej dowodów, że wyszukiwania w Google'u pokazują inny obraz świata niż ten, który widzimy na co dzień? Weźmy wyszukiwania związane z trudnym pytaniem o to, czy mieć lub nie mieć dzieci. Niektórzy z nas boją się, że mogą podjąć w tej kwestii niewłaściwą decyzję, przy czym prawie zawsze chodzi tutaj o to, czy nie będziemy żałować, jeśli postanowimy nie zostać rodzicami. Internauci siedmiokrotnie częściej wpisują w wyszukiwarkę pytanie, czy będą żałować tego, że nie mają dzieci, niż czy będą żałować, jeśli zdecydują się jednak mieć potomstwo.

Ale kiedy podejmą już decyzję – żeby się rozmnożyć albo nie (ewentualnie zaadoptować dziecko) – przyznają czasem, że żałują dokonanego wyboru. Może to być nieco szokujące, lecz wyżej podane proporcje zmieniają się post factum. Dorośli obarczeni dziećmi trzy i sześć dziesiątych razy częściej niż osoby bezdzietne informują wyszukiwarkę, że podjęli niewłaściwą decyzję[10].

W bieżącym rozdziale powinniśmy jednak mieć w pamięci jedno zastrzeżenie: Google częściej pokazuje nam raczej niestosowne myśli użytkowników, takie, którymi, jak sądzą, nie powinni się dzielić z nikim innym. Jeżeli chcemy zatem ujawnić myśli na ogół skrywane przez internautów, wyszukiwarka okaże się użyteczna. A spora rozbieżność

między liczbą osób, które żałują tego, że mają lub nie mają dzieci, mówi nam, że tego rodzaju „niestosowne myśli" często się ludziom zdarzają.

Przerwijmy na chwilę i zastanówmy się, co to w ogóle znaczy, że wpisujemy w wyszukiwarkę takie zdanie, jak: „Żałuję, że mam dzieci". Google wydaje się źródłem, w którym możemy szukać informacji bezpośrednich, na przykład o pogodzie, o tym, kto wygrał wczorajszy mecz albo kiedy wybudowano Statuę Wolności. Czasem jednak wpisujemy w wyszukiwarkę swoje nieocenzurowane przemyślenia bez większej nadziei na to, że znajdziemy pomoc. W takich wypadkach Google staje się czymś w rodzaju konfesjonału.

Każdego roku w sieci pojawiają się tysiące wpisów w rodzaju: „Nie znoszę, jak jest zimno", „Ludzie mnie irytują" albo „Jest mi smutno". Rzecz jasna, autorami tysięcy wpisów typu „Jest mi smutno" zaledwie drobny ułamek setek milionów osób, którym w danym roku doskwiera uczucie smutku. Z moich badań wynika, że wpisów zawierających jakieś przemyślenia, a nie poszukujących informacji, dokonuje jedynie niewielka grupa internautów spośród wszystkich tych, którzy podobnie myślą albo znajdują się w analogicznym stanie ducha. Moje badania wskazują zatem również, że siedem tysięcy wpisów w rodzaju „Żałuję, że mam dzieci", jakie co roku Amerykanie powierzają sieci, pochodzi tylko od niewielkiej części mieszkańców USA, którym taka myśl w ogóle przyszła do głowy.

Rodzicielstwo z pewnością dostarcza mnóstwa radości wielu ludziom, może nawet większej części populacji. A pomimo obaw mojej mamy, że „ty i ta twoja głupia analiza danych" ograniczą liczbę jej wnuków, badania wcale nie zniechęciły mnie do posiadania dzieci. Ale też ta „nieprzyzwoita" myśl – że lepiej byłoby nie zostać ojcem lub matką – jest interesująca i pokazuje kolejny aspekt człowieczeństwa, którego nie zobaczymy w tradycyjnych zbiorach danych. Kultura nieustannie zalewa nas obrazkami cudownych, szczęśliwych rodzin, a większość rodziców nigdy nawet nie sądziłaby, że może kiedyś żałować decyzji o posiadaniu dzieci. Ale niektórzy żałują. I zwykle nie zwierzają się z tego nikomu – z wyjątkiem Google'a.

PRAWDA O SEKSIE

Jaki jest odsetek gejów w naszej populacji? To pytanie obrosło już legendą pośród badaczy ludzkiej seksualności i stanowi jeden z najtwardszych orzechów do zgryzienia dla przedstawicieli nauk społecznych. Dzisiejsi psychologowie nie wierzą w słynne wyliczenia Alfreda Kinseya – oparte na sondażach zbyt często przeprowadzanych wśród więźniów i prostytutek – że geje stanowią około dziesięciu procent ludności. Reprezentatywne badania współczesne wskazują, że liczba ta wynosi od dwóch do trzech procent. Ale odpowiadając na pytania o swoją orientację seksualną, ludzie kłamali od dawna. Wydaje mi się, że potrafię wykorzystać big data, by podać wynik bliższy prawdy niż kiedykolwiek wcześniej.

Przypatrzmy się najpierw danym sondażowym. Ankiety powiadają, że w stanach tolerancyjnych mieszka znacznie więcej gejów niż w nietolerancyjnych. Na przykład według badań Gallupa populacja gejów jest proporcjonalnie dwukrotnie większa w Rhode Island, stanie, w którym poparcie dla małżeństw homoseksualnych jest najwyższe[11] w kraju, niż w Missisipi, gdzie jest ono z kolei najmniejsze.

Istnieją dwa możliwe wytłumaczenia tego faktu. Pierwsze: geje urodzeni w stanach nietolerancyjnych przeprowadzają się z czasem do stanów tolerancyjnych. Drugie – mężczyźni w stanach nietolerancyjnych mogą nie przyznawać się do swoich skłonności homoseksualnych, a pytani o nie wprost, najprawdopodobniej kłamią.

Wyjaśnienie numer jeden – mobilność gejów – możemy naświetlić lepiej, sięgając do źródła informacji pochodzących z grupy big data, czyli w tym wypadku z Facebooka, który pozwala swoim użytkownikom wskazywać preferencje płciowe. I tak około dwóch i pół procent mężczyzn korzystających z tego portalu społecznościowego twierdzi, że interesuje się przedstawicielami swojej płci – co zgadza się mniej więcej z wynikami sondaży. Facebook potwierdza też spore różnice w populacji gejowskiej w stanach o wysokiej i niskiej tolerancji: w Rhode Island żyje dwa razy więcej gejów niż w Missisipi.

Facebook dostarcza również informacji o tym, jak ludzie się przemieszczają. Udało mi się na przykład ustalić, z jakich miast pochodzi reprezentatywna grupa jawnie homoseksualnych użytkowników Facebooka, dzięki czemu mogłem oszacować, ilu gejów przeprowadza się ze stanów nietolerancyjnych w bardziej liberalne obyczajowo rejony kraju. Odpowiedź? Przypadki takich przenosin zdarzają się dość często – na przykład z Oklahoma City do San Francisco. Wydaje mi się jednak, że liczba mężczyzn, którzy pakują swoje kompakty Judy Garland i przenoszą się w jakieś bardziej permisywne miejsce, wyjaśnia bardzo niewiele, jeżeli chodzi o różnicę w wielkości populacji gejowskiej w stanach tolerancyjnych i nietolerancyjnych[9*].

Poza tym dzięki Facebookowi możemy przyjrzeć się licealistom. Jest to specyficzna grupa społeczna, bo chłopcy w wieku szkolnym rzadko mogą sami decydować o tym, gdzie mieszkają. Jeśli mobilność wyjaśnia różnice populacyjne wśród jawnie zdeklarowanych homoseksualistów, to nie powinny one występować między uczniami szkół średnich. A co mówią dane? Że w nietolerancyjnych stanach USA zdeklarowanych gejów wśród młodzieży szkolnej znajdziemy daleko mniej. W Missisipi na tysiąc uczniów do orientacji homoseksualnej otwarcie przyznaje się zaledwie dwóch – mobilność więc to nie wszystko.

Jeśli w każdym stanie przychodzi na świat podobny odsetek gejów, a mobilność nie tłumaczy w pełni, dlaczego w niektórych stanach mieszka ich tak wielu, to znaczy, że spora liczba homoseksualistów ukrywa swoją orientację – co prowadzi nas z powrotem do wyszukiwarki Google'a, której, jak się okazuje, dużo ludzi chętnie powierza różne informacje na swój temat.

Czy istnieje taki sposób wykorzystania wpisów pornograficznych, by można było sprawdzić, ilu gejów naprawdę mieszka w różnych stanach? Owszem. Wedle mojej oceny – dokonanej na podstawie wyszukiwań w Google'u i w Google AdWords – w Stanach Zjednoczonych pornografii gejowskiej szuka około pięciu procent męskiej populacji (uwzględniam tu z jednej strony takie wpisy jak „Rocket Tube", czyli popularny gejowski portal pornograficzny, a z drugiej „gejowskie porno").

Czy te wyniki są różne w różnych częściach kraju? Na ogół więcej wyszukiwań gejowskiej pornografii znajdziemy w stanach tolerancyjnych, co jest logiczne, biorąc pod uwagę, że niektórzy geje przeprowadzają się z miejsc mniej tolerancyjnych do bardziej liberalnych obyczajowo. Ale różnice wcale nie są tak duże, jak sugerują sondaże oraz Facebook. Ustaliłem na przykład, że w Missisipi gejowskiego porno szuka w sieci cztery i osiem dziesiątych procent mężczyzn, czyli znacznie więcej, niż wynikałoby z tradycyjnych ankiet czy badań Facebooka, i niewiele mniej niż pięć i dwie dziesiąte procent męskich mieszkańców Rhode Island.

A zatem jaki odsetek Amerykanów to geje? Prawdopodobnie około pięciu procent, czyli tyle, ile mamy homoseksualnych wyszukiwań w sieci[12]. Ale istnieje jeszcze jedna, mniej bezpośrednia metoda osiągnięcia tego samego wyniku, wymaga ona jednak szczypty analityki. Możemy bowiem wykorzystać związek między tolerancją a liczbą zdeklarowanych gejów w danym regionie czy mieście. Proszę was o chwilę cierpliwości.

Z moich wstępnych badań wynika, że w danym stanie każde dwadzieścia punktów procentowych poparcia dla małżeństw homoseksualnych oznacza, że około półtora raza więcej mężczyzn otwarcie zadeklaruje swoją homoseksualną orientację na Facebooku. Wychodząc od tych danych, można obliczyć w przybliżeniu, ilu mężczyzn urodzonych w hipotetycznie całkowicie tolerancyjnym miejscu, gdzie dajmy na to sto procent mieszkańców poparło małżeństwa homoseksualne, przyzna bez ogródek, że są gejami. Moje szacunki wskazują, że będzie to około pięciu procent, co bardzo ładnie zgadza się z danymi uzyskanymi na podstawie wyszukiwań gejowskiej pornografii. A najbardziej tolerancyjne dla gejów środowisko stanowią licealiści z Bay Area w Kalifornii – mniej więcej cztery procent z nich przyznaje otwarcie na Facebooku, że są gejami[13]. I ten wynik również zdaje się potwierdzać moje wcześniejsze obliczenia.

Powinienem tu dodać, że nie udało mi się na razie zebrać podobnych informacji na temat homoseksualnych kobiet. Dane z wyszukiwań pornograficznych są w ich wypadku mniej użyteczne, ponieważ znacznie mniej kobiet ogląda pornografię, liczby są tu zatem mniej

reprezentatywne. Tym bardziej że wśród oglądających treści pornograficzne pań nawet te z nich, które w prawdziwym życiu odczuwają głównie pociąg do mężczyzn, najwyraźniej chętnie wyszukują pornografię lesbijską. Aż dwadzieścia procent filmów oglądanych przez kobiety na PornHubie to pornografia lesbijska.

Rzecz jasna, owe pięć procent gejów wśród mężczyzn to jedynie wynik przybliżony. Istnieją przecież także biseksualiści – a niektórzy faceci, zwłaszcza w młodym wieku, nie są pewni swojej orientacji. Jest więc absolutnie oczywiste, że nie da się policzyć gejów tak dokładnie, jak głosujących wyborców albo widzów filmowych.

Ale jeśli chodzi o wyniki moich badań, to jedno staje się jasne: bardzo wielu mężczyzn w USA, szczególnie w stanach nietolerancyjnych, nadal ukrywa swoją orientację seksualną. Kryptogeje nie ujawniają się na Facebooku ani w sondażach, a w wielu wypadkach bywają nawet żonaci.

Okazuje się, że kobiety nierzadko podejrzewają swoich mężów o homoseksualizm, co widać w zaskakująco często pojawiającym się autouzupełnieniu: „Czy mój mąż… jest gejem?". Jako uzupełnienie wyszukiwań zaczynających się od frazy „Czy mój mąż…" „gej" występuje o dziesięć procent częściej niż zajmująca drugie miejsce „zdrada" – „Czy mój mąż… mnie zdradza?". Pojawia się też osiem razy częściej niż „jest alkoholikiem" i dziesięć razy częściej niż „cierpi na depresję".

Najbardziej wymowne okazuje się być może to, że wyszukiwania dotyczące wątpliwości co do orientacji seksualnej mężów prowadzą głównie kobiety z najmniej tolerancyjnych części USA, największy odsetek takich internautek znajdziemy bowiem w Karolinie Południowej i Luizjanie. Co więcej, w dwudziestu jeden z dwudziestu pięciu stanów, gdzie to pytanie zadaje się najczęściej, poparcie dla małżeństw homoseksualnych jest niższe niż średnia krajowa.

Google i portale pornograficzne to nie jedyne użyteczne źródła informacji na temat męskiej seksualności. Big data dostarcza nam więcej materiałów świadczących o tym, że wielu panów ukrywa swoją orientację seksualną przed światem. Przeanalizowałem na przykład ogłoszenia na Craigslist, zamieszczane przez mężczyzn poszukujących partnerek lub partnerów do „niezobowiązujących spotkań". Odsetek ogłoszeń, których autorzy poszukują mężczyzn, jest wyższy w mniej

tolerancyjnych stanach, a na pierwszych miejscach plasują się wśród nich Kentucky, Luizjana i Alabama.

By zajrzeć jeszcze dalej za zasłonę milczenia, powróćmy do danych z wyszukiwania w Google'u i bądźmy nieco bardziej szczegółowi. Otóż jednym z najczęstszych wpisów pojawiających się bezpośrednio przed albo po „gejowskim porno" jest „test gejowski" (testy takie mogą rzekomo rozstrzygnąć, czy wypełniający je mężczyźni są homoseksualistami). A wyszukiwania „testów gejowskich" zdarzają się dwa razy częściej w najmniej tolerancyjnych stanach.

Co znaczy, że mężczyzna na przemian poszukuje „gejowskiego porno" i „testu gejowskiego"? Zapewne to, że nie jest pewien swojej orientacji i może nawet w duchu zadręcza się tą niepewnością. Można chyba zasadnie podejrzewać, że niektórzy panowie mają nadzieję znaleźć w takich testach potwierdzenie, że zainteresowanie gejowską pornografią nie czyni ich gejami.

Baza danych Google'a nie pozwala śledzić historii wyszukiwań konkretnego użytkownika – ale w 2006 roku serwis AOL przekazał próbkę wpisów swoich użytkowników badaczom akademickim. Oto przykład erotycznych odsłon anonimowego internauty, wybranych z sześciu dni, od piątku do środy włącznie:

piątek 03:49:55	darmowe fotki gejowskie
piątek 03:59:37	fotki gejów z szatni
piątek 04:00:14	fotki gejowskie
piątek 04:00:35	fotki stosunków gejowskich
piątek 05:08:23	długi quiz gejowski
piątek 05:10:00	dobry test gejowski
piątek 05:25:07	testy gejowskie dla zdezorientowanych
piątek 05:26:38	testy gejowskie
piątek 05:27:22	testy czy jestem gejem
piątek 05:29:18	fotki gejowskie
piątek 05:30:01	fotki nagich mężczyzn
piątek 05:32:27	darmowe fotki nagich mężczyzn
piątek 05:38:19	gorące fotki gejowskie

piątek 05:41:34	gorący męski seks analny
środa 13:37:37	testy czy jestem gejem
środa 13:41:20	testy gejowskie
środa 13:47:49	gorący męski seks analny
środa 13:50:31	darmowe filmy gejowskie

Te dane niewątpliwie wskazują na człowieka, który nie czuje się dobrze ze swoją seksualnością. A na podstawie materiałów z Google'a możemy przypuszczać, że w Ameryce żyje wielu takich mężczyzn jak on. Nawiasem mówiąc, większość z nich mieszka w stanach mniej tolerancyjnych wobec związków osób tej samej płci.

Aby jeszcze lepiej przyjrzeć się prawdziwym ludziom stojącym za tymi wszystkimi liczbami, zapytałem pewnego psychiatrę z Missisipi, specjalizującego się w pomocy gejom, którzy ukrywają swoją orientację przed otoczeniem, czy między jego pacjentami znalazłby się ktoś, kto zechciałby ze mną porozmawiać. Na ten mój apel odpowiedział tylko jeden. Powiedział, że jest profesorem na emeryturze, przekroczył sześćdziesiąty rok życia i od ponad czterdziestu lat żyje w związku małżeńskim. Dopiero dziesięć lat przed naszym spotkaniem poszedł do psychiatry pod wpływem przemożnego stresu i w końcu przyznał się do swojej orientacji. Podobno zawsze zdawał sobie sprawę, że pociągają go mężczyźni, sądził jednak, że to powszechne zjawisko, coś, co wszyscy faceci po prostu ukrywają przed światem. Wkrótce po rozpoczęciu terapii odbył swój pierwszy i jedyny gejowski stosunek seksualny – ze zbliżającym się do trzydziestki studentem – i uznał to doświadczenie za „cudowne".

Nie utrzymuje natomiast stosunków seksualnych z żoną. Twierdzi, że czułby się winny, gdyby rozwiązał małżeństwo albo otwarcie zaczął się spotykać z jakimś mężczyzną. I że żałuje właściwie każdej swojej ważniejszej decyzji życiowej.

Emerytowany profesor i jego żona spędzą więc kolejną noc bez romantycznej miłości i bez seksu. Pomimo ogromnego postępu na tym polu, utrzymująca się nietolerancja zmusi miliony innych Amerykanów do podobnych zachowań.

Być może nie szokuje was wiadomość, że pięć procent amerykańskich mężczyzn to geje, a wielu z nich ukrywa swoją orientację seksualną przed otoczeniem, choć były takie czasy, kiedy większość naszego społeczeństwa poczułaby się w obliczu tych faktów wstrząśnięta. Na świecie jednak istnieją jeszcze miejsca, gdzie podobna informacja wywołałaby szok również dzisiaj.

„U nas, w Iranie, nie ma homoseksualistów, jak u was[14] – upierał się w roku 2007 Mahmud Ahmadineżad, prezydent Iranu. – My nie znamy tu takiego zjawiska". W podobnym duchu wypowiadał się Anatolij Pachomow, burmistrz rosyjskiego miasta Soczi na krótko przed otwarciem Zimowych Igrzysk Olimpijskich w roku 2014: „W naszym mieście nie ma gejów"[15]. Aktywność internetowa wskazuje jednak na to, że zarówno w Soczi, jak i w Iranie równie wiele osób interesuje się pornografią gejowską[16].

Każe nam to postawić dość oczywiste pytanie: czy w Stanach Zjednoczonych znajdziemy jeszcze jakąś sferę powszechnych zainteresowań seksualnych, które wciąż byłyby uznawane za szokujące?

Wszystko zależy od tego, co uważamy za powszechne i jak łatwo jest nas zaszokować.

Większość najczęstszych wpisów na portalu PornHub nie wzbudzi naszego zdziwienia – znajdziemy wśród nich takie słowa jak „nastolatki", „trójkąt" i „obciąganie", jeżeli chodzi o mężczyzn, i takie frazy jak „namiętny seks", „ssanie piersi" oraz „facet robiący minetę", jeżeli chodzi o kobiety.

Gdy odejdziemy jednak od zainteresowań mainstreamowych, okaże się, że PornHub mówi nam o wielu fetyszach i zachciankach, o których istnieniu mogliście nie mieć pojęcia. Są więc na przykład kobiety, które poszukują „jabłek w odbycie" albo filmów przedstawiających „ruchanie wypchanych zwierząt", i są mężczyźni poszukujący „fetyszy smarkowych" albo pragnący „ukrzyżowania nago". Takie wyszukiwania zdarzają się jednak rzadko, mniej więcej zaledwie dziesięć razy na miesiąc, nawet na tak wielkim portalu pornograficznym jak PornHub.

Kolejny wniosek, który wydaje się całkowicie jasny, kiedy przyjrzymy się danym pornhubowym, brzmi, że zawsze gdzieś na świecie jest nasza druga połowa. Nie ma nic dziwnego w tym, że kobiety często szukają „wysokich", „ciemnowłosych" i „przystojnych" facetów – czasami jednak szukają też facetów „niskich", „bladych" i „brzydkich". Panie niekiedy pragną również mężczyzn „niepełnosprawnych", „korpulentnego gościa z małym fiutem" albo „grubego brzydkiego starucha". Z kolei panowie często wyszukują kobiety „szczupłe", „z dużymi cyckami" albo „blondynki", choć zdarza im się szukać także „grubasek", kobiet „z malutkimi cycuszkami" albo „o zielonych włosach". Niektórzy chcą kobiet „łysych", „karlic" lub „pozbawionych brodawek piersiowych". Dane te powinny brzmieć krzepiąco dla wszystkich facetów, którzy nie są wysocy, ciemnowłosi i przystojni, oraz dla wszystkich kobiet, które nie są szczupłe, nie mają dużego biustu i nie są blondynkami[10*].

A inne wyszukiwania, które można uznać zarówno za częste, jak i zaskakujące? Jeżeli chodzi o mnie, to wśród stu pięćdziesięciu najczęstszych wyszukiwań ze strony mężczyzn najbardziej zaskoczyły mnie te kazirodcze wpisy, które omawiałem w rozdziale freudowskim. Inne rzadko dyskutowane obiekty męskiego pożądania to „hermafrodyci" (na siedemdziesiątym siódmym miejscu) i „babcia" (miejsce sto dziesiąte). W sumie mniej więcej jeden i cztery dziesiąte procent mężczyzn szuka na PornHubie kobiet z penisami. Jakieś sześć dziesiątych procent (cztery dziesiąte dla mężczyzn poniżej trzydziestego czwartego roku życia) wyszukiwań dotyczy kobiet w podeszłym wieku. Natomiast zaledwie jeden na dwadzieścia cztery tysiące wpisów wyraźnie odnosi się do małych dzieci, których nie można jeszcze zaliczyć do nastolatków – może to być jednak związane z tym, że PornHub z oczywistych względów banuje wszystkie formy pornografii dziecięcej, a jej posiadanie jest w USA nielegalne.

Z kolei wśród najczęściej spotykanych wpisów ze strony kobiet znajdziemy taki rodzaj pornografii, który, muszę was ostrzec, zbulwersuje wielu czytelników – chodzi bowiem o seks połączony z przemocą wobec kobiet. Aż dwadzieścia pięć procent wpisów pań poszukujących pornografii heteroseksualnej eksponuje ból i/albo poniżenie – na przykład „bolesny seks analny z płaczem", „publiczne

pohańbienie" czy „ekstremalnie brutalny seks zbiorowy". Pięć procent internautek szuka scen seksu, do którego dochodzi bez zgody kobiety – scen „gwałtu" albo „seksu wymuszonego" – mimo że takie filmy są zabronione na PornHubie. Okazuje się, że pornografii tego typu kobiety poszukują dwukrotnie częściej niż mężczyźni. Według moich badań wszystkie podgatunki pornograficzne, w których dochodzi do aktów przemocy seksualnej wobec kobiet, są nieproporcjonalnie częściej preferowane przez panie.

Oczywiście kiedy próbujemy pogodzić się z tym faktem, z pewnością należy pamiętać, że istnieje zasadnicza różnica pomiędzy fantazjami i prawdziwym życiem. Owszem, wśród mniejszości odwiedzających PornHub kobiet znajdziemy ich pewien margines, poszukujący – bezskutecznie – scen gwałtu. Jest rzeczą oczywistą, że nie oznacza to, że te kobiety chcą być gwałcone naprawdę, a już na pewno nie czyni to gwałtu mniej okropną zbrodnią. Dane pornograficzne mówią nam tylko tyle, że ludzie miewają czasem takie fantazje, których woleliby nie mieć i z których być może nikomu się nie zwierzają.

Nasze dusze i głowy nie są wyłącznie składnicami fantazji. Kryjemy też wiele innych tajemnic seksualnych – na przykład w kwestii częstotliwości stosunków płciowych[17].

We wstępie zauważyłem, że według sondaży Amerykanie zużywają o wiele więcej prezerwatyw, niż wynosi ich całoroczna sprzedaż w USA. Można by zatem pomyśleć, że w gruncie rzeczy respondenci mówią tylko tyle, że zużywają więcej gumek niż w rzeczywistości. Mamy jednak dane, które wskazują, że wyolbrzymiamy również liczbę swoich stosunków seksualnych – i od tego chyba trzeba zacząć. Mniej więcej jedenaście procent[18] kobiet pomiędzy piętnastym i czterdziestym czwartym rokiem życia twierdzi, że prowadzi aktywne życie seksualne, nie jest w ciąży i nie używa żadnych środków antykoncepcyjnych. Przyjmując w tej sytuacji nawet dość konserwatywne założenia co do liczby ich corocznych stosunków płciowych, uczeni szacują, że co najmniej dziesięć procent[19] z tych pań powinno zachodzić w ciążę każdego miesiąca. Gdyby tak jednak było, to liczba zapłodnień wśród

tych respondentek przekroczyłaby liczbę wszystkich ciąż w Stanach Zjednoczonych (w ciążę w USA zachodzi jedna na sto trzynaście kobiet[20] w wieku rozrodczym). Ale w naszej opętanej seksem kulturze być może trudno jest przyznać, że po prostu wcale nie kochamy się zbyt często.

Jeżeli jednak szukamy zrozumienia lub rady, to i w kwestii seksu jesteśmy skłonni zwierzyć się wyszukiwarce. W Google'u znajdziemy szesnaście razy więcej skarg na współmałżonka, który nie chce z nami współżyć, niż na współmałżonka, który nie chce z nami rozmawiać, oraz pięć i pół razy więcej skarg na partnera stanu wolnego odmawiającego nam seksu niż na partnera nieodpisującego na nasze esemesy.

Wpisy w Google'u wskazują również nieoczekiwanych winowajców w wielu takich pozbawionych seksu związkach – bo o wstrzemięźliwość seksualną dwukrotnie częściej oskarżani są partnerzy. Zdecydowanie najliczniej występujący w sieci zarzut wobec panów brzmi: „Mój facet nie chce uprawiać ze mną seksu" (wyszukiwania w Google'u nie są różnicowane ze względu na płeć, ale ponieważ z poprzedniej analizy wynika, że dziewięćdziesiąt pięć procent mężczyzn jest heteroseksualnych, możemy się domyślać, że tylko bardzo niewiele takich wpisów na temat partnerów pochodzi od mężczyzn).

Jak powinniśmy interpretować te dane? Czy chłopcy i mężczyźni naprawdę częściej powstrzymują się od seksu niż dziewczęta i kobiety? Niekoniecznie. Jak wspominaliśmy wcześniej, wyszukiwania w Google'u mogą dotyczyć zwłaszcza takich spraw, o których wstydzimy się mówić otwarcie. Być może mężczyznom łatwiej przychodzi przyznać przed znajomymi, że ich partnerek nie interesuje seks. Ale nawet jeśli z danych Google'a nie wynika, że naprawdę unikamy seksu dwukrotnie częściej niż panie, to można jednak założyć, że generalnie facetom zdarza się to częściej, niż przypuszczamy.

Dane z Google'a sugerują także powód, dla którego prawdopodobnie Amerykanie i Amerykanki tak często unikają seksu – jest to olbrzymi niepokój albo lęk, w większości zupełnie nieuzasadniony. Zacznijmy od męskich obaw. To nic nowego, że mężczyźni martwią się tym, jak

zostali obdarzeni przez naturę, ważniejsza wydaje się jednak skala tych fobii.

Osobnicy mojej płci zadają w sieci więcej pytań o swój narząd seksualny niż o jakąkolwiek inną część ciała: więcej niż o płuca, wątrobę, stopy, uszy, nos, szyję i mózg razem wzięte. Częściej szukają odpowiedzi na pytanie o to, jak powiększyć penisa, niż nastroić gitarę, zrobić omlet albo zmienić koło w samochodzie. Jeżeli biorą sterydy, martwią się głównie tym, czy wskutek ich zażywania nie zmniejszy się rozmiar ich członka. Podobnie jeśli chodzi o zmiany fizyczne i umysłowe powiązane z procesem starzenia się – faceci przede wszystkim boją się, czy ich penis z wiekiem się nie zmniejszy.

Uwaga na marginesie: jednym z najczęściej zadawanych pytań w Google'u dotyczących męskich genitaliów jest: „Jak dużego mam członka?". To, że mężczyźni odpowiedzi na to pytanie poszukują w sieci, zamiast sięgnąć po miarkę, wydaje mi się świetnie oddawać ducha naszej epoki cyfrowej[11*].

A czy kobiety interesują się rozmiarami penisów? Wyszukiwania w Google'u pokazują, że rzadko, a dokładnie sto siedemdziesiąt razy rzadziej niż mężczyźni. Owszem, jeśli już panie niepokoi rozmiar penisa partnera, zazwyczaj ma to związek z jego wielkością, ale wcale niekoniecznie chodzi o to, że jest za mały – uskarżając się na rozmiar członka, kobiety w czterdziestu procentach przypadków piszą, że jest za duży. Najczęstszym uzupełnieniem frazy „w czasie seksu zdarza mi się..." jest dla nich „ból" (pierwszą piątkę dopełniają „krwawić", „sikać", „płakać" i „puszczać bąki"). Mimo to wśród mężczyzn pytających, jak zmienić wielkość penisa, zaledwie jeden procent poszukuje sposobów jego zmniejszenia.

Drugie najczęściej zadawane przez panów pytanie z dziedziny seksu brzmi: „Jak przedłużyć stosunek seksualny?". I tym razem męskie lęki nie współgrają z obawami kobiet, które o to, jak szybciej doprowadzić chłopaka do orgazmu, pytają mniej więcej z taką samą częstotliwością jak o to, w jaki sposób opóźnić jego wytrysk. Najczęściej natomiast interesują się nie tym, jak szybko facet dochodzi, lecz czy dochodzi w ogóle.

My, mężczyźni, nie rozmawiamy zbyt często na temat naszego wyglądu, ale chociaż jest prawdą, że generalnie rzecz biorąc, zagadnienie aparycji bardziej zajmuje kobiety, to wcale nie aż tak przytłaczająco częściej, jak wskazywałyby na to stereotypy. Według mojej analizy danych z Google AdWords, gdzie możemy się dowiedzieć, jak często ludzie odwiedzają różne portale, od mężczyzn pochodzi czterdzieści dwa procent wpisów dotyczących urody i sprawności fizycznej, odchudzania – trzydzieści trzy procent, a chirurgii plastycznej – trzydzieści dziewięć. Wśród wszystkich wyszukiwań dotyczących torsu, zaczynających się od „W jaki sposób…", około dwudziestu procent stanowią pytania o to, jak pozbyć się za dużych męskich piersi.

Ale nawet jeśli liczba mężczyzn, którzy nie są zbyt zadowoleni ze swojego wyglądu, jest wyższa, niż wydaje się większości z nas, to kobiety zdecydowanie ich wyprzedzają w tej kwestii. Co mówi na ten temat cyfrowe serum prawdy? Że w Stanach Zjednoczonych co roku ponad siedem milionów razy wyszukuje się w sieci informacji dotyczących implantów biustu, a według oficjalnych statystyk operacji ich wszczepienia poddaje się rocznie około trzystu tysięcy kobiet.

Kobiety niezwykle przejmują się też wyglądem swoich pośladków, chociaż w ostatnich czasach większa część z nich całkowicie zmieniła swoje poglądy, jeżeli chodzi o to, co im się w nich może nie podobać.

Jeszcze w 2004 roku przynajmniej w niektórych regionach Stanów Zjednoczonych internautki najczęściej szukały odpowiedzi na pytanie, jak zmniejszyć swoją pupę, zamiar jej powiększenia deklarowały natomiast głównie mieszkanki tych okolic, gdzie przeważała liczebnie ludność czarnoskóra. Ale poczynając od roku 2010, pragnienie posiadania większych pośladków ogarnęło już całe Stany Zjednoczone. Zainteresowanie nimi wzrosło trzykrotnie, choć chyba nie stały się przez to trzykrotnie większe. W 2014 roku już w każdym stanie liczba pytań w sieci o to, jak powiększyć swój tyłek, przekraczała liczbę pytań o to, jak go zmniejszyć. Obecnie na każde pięć wpisów dotyczących implantów biustu pojawia się minimum jeden związany z implantami pośladków (dziękujemy ci, Kim Kardashian!).

Czy rosnące zainteresowanie większą pupą u pań przekłada się na zainteresowanie nią u mężczyzn? Co ciekawe, tak. Wyszukiwania treści pornograficznych obejmujących „duże tyłki", a preferowanych niegdyś głównie w społecznościach afroamerykańskich, zaczęły się ostatnio cieszyć coraz większą popularnością w całych Stanach Zjednoczonych.

A czego jeszcze mężczyźni szukają w kobiecym ciele? Jak pisałem wcześniej – i co dla wielu z nas jest porażająco oczywiste – lubią duże piersi. Mniej więcej dwanaście procent nietypowych wyszukiwań na portalach pornograficznych dotyczy obfitych kształtów – wyszukiwania małych piersi zdarzają się dwadzieścia razy rzadziej.

Nie dowodzi to natomiast, że wszyscy mężczyźni lubią implanty u pań. Mniej więcej trzy procent facetów szukających dużych piersi w Internecie stwierdza przy tym wyraźnie, że chcą oglądać tylko naturalne biusty.

Z kolei wpisy dotyczące żon i implantów piersi dzielą się równo – w połowie są to pytania o to, jak skłonić małżonkę, by powiększyła biust, a w połowie wyrażają zdziwienie, że kobieta chce się poddać takiej operacji.

Spójrzmy też na najczęściej spotykany wpis na temat piersi partnerki: „Uwielbiam cycki mojej dziewczyny". Trudno powiedzieć, jakie informacje próbują znaleźć mężczyźni wklepujący taki tekst w wyszukiwarkę.

Kobiety również zadają w sieci pytania dotyczące swoich genitaliów, a na temat pochwy wpisują ich nawet niemal tyle samo, ile faceci na temat penisa. Kobiecy lęk o pochwę często ma związek ze zdrowiem, ale przynajmniej trzydzieści procent wyszukiwań dotyka innych problemów. Panie chcą między innymi wiedzieć, jak golić włosy łonowe, uczynić pochwę ciaśniejszą i nadać jej przyjemniejszy smak. Jak wspomniałem już wcześniej, martwią się też powszechnie jej zapachem i pytają, co zrobić, by ładniej pachniała.

Zauważają najczęściej, że ich pochwy wydzielają woń ryb, a potem kolejno octu, amoniaku, czosnku, sera, niemytego ciała, uryny, chleba, wybielacza, odchodów, potu, metalu, spoconych nóg, śmieci i zgniłego mięsa.

Mężczyźni w Internecie na ogół nie zadają zbyt wielu pytań dotyczących żeńskich genitaliów. Robią to mniej więcej równie często jak kobiety, które pytają o penisa partnera.

Pisząc w sieci o waginie partnerki, panowie zwykle skarżą się na to, o co kobiety martwią się najbardziej – na jej niemiły zapach. Na ogół jednak starają się znaleźć odpowiedź na pytanie, jak powiedzieć kobiecie, że nieładnie pachnie, i nie sprawić jej przy tym przykrości. Czasem zapachowe wpisy mężczyzn ujawniają z kolei ich własne obawy, na przykład kiedy pytają, czy wagina partnerki pachnąca, powiedzmy, kondomami albo nasieniem innego mężczyzny, nie jest dowodem zdrady.

Do jakich wniosków prowadzą te wszystkie utajone lęki? Tu najwyraźniej mam dla was dobre wieści. Google daje uzasadnione powody, żebyśmy martwili się tymi sprawami mniej niż do tej pory. Większość najskrytszych ludzkich obaw o to, jak postrzegają nas partnerzy seksualni, wydaje się bezpodstawna. W samotności, przy swoich komputerach, nie mając żadnej motywacji, by kłamać, okazujemy się niepowierzchowni i tolerancyjni. Prawda wygląda tak, że jesteśmy zbyt zajęci oceną własnego wyglądu i już nie wystarcza nam energii na krytyczną ocenę aparycji naszych partnerów.

Istnieje też prawdopodobnie związek pomiędzy dwoma dużymi problemami, jakie ujawniają wyszukiwania w Google'u: brakiem seksu i niepewnością co do swojej atrakcyjności i swoich umiejętności seksualnych. Wygląda na to, że te kwestie są ze sobą blisko powiązane. Może gdybyśmy mniej się przejmowali seksem, kochalibyśmy się częściej.

Co jeszcze Google może nam powiedzieć o seksie? Przeprowadźmy na przykład pojedynek płci i przyglądając się wpisom osób, które szukają sposobów na udoskonalenie swoich umiejętności w seksie oralnym, sprawdźmy, kto jest w tym względzie bardziej wielkoduszny, mężczyźni czy kobiety[21]. Która płeć częściej szuka porad w tej kwestii? Kto jest hojniejszy w łóżku, panowie czy panie? Tak, panie. Sumując wszystkie dane, oceniam, że kobiety dwa razy częściej niż mężczyźni proszą o wskazówki, dzięki którym mogłyby sprawiać większą rozkosz partnerowi.

Panowie szukający podpowiedzi, jak uprawiać miłość francuską, na ogół wcale nie robią tego po to, by sprawić przyjemność partnerce. Owszem, pytają czasem, w jaki sposób doprowadzić kobietę do orgazmu, lecz równie często chcą wiedzieć, jak mogliby zaspokoić oralnie samych siebie – i to jest jedna z moich ulubionych informacji ze zbioru danych Google'a.

PRAWDA O NIENAWIŚCI I UPRZEDZENIACH

Seks i miłość to oczywiście nie jedyne wstydliwe dla nas tematy, a zatem nie tylko w tych sprawach ukrywamy prywatne myśli przed światem. Wielu ludzi ze zrozumiałych powodów nie przyznaje się otwarcie do swoich uprzedzeń. Być może da się powiedzieć, że osiągnęliśmy w tej kwestii postęp, ponieważ wiele osób obawia się dziś krytyki, jeżeli przyznaje publicznie, że ocenia innych według pochodzenia etnicznego, orientacji seksualnej czy wyznania – chociaż pokaźna liczba Amerykanów wciąż właśnie tak postępuje (muszę przestrzec tutaj czytelników, że ten podrozdział również zawiera bulwersujący materiał).

Widać to dobrze w Google'u, gdzie internauci zadają czasem pytania w rodzaju „Dlaczego czarni są chamscy?" albo „Dlaczego Żydzi to z natury źli ludzie?". Poniżej podaję najczęstsze pejoratywne określenia pojawiające się w wyszukiwaniach na temat określonych grup społecznych, po pięć dla każdej z nich[22].

	1	2	3	4	5
Afroamerykanie	chamy	rasiści	głupcy	brzydale	lenie
Żydzi	źli	rasiści	brzydale	skąpcy	chciwcy
muzułmanie	źli	terroryści	okrutnicy	skłonni do przemocy	niebezpieczni
Meksykanie	rasiści	głupcy	brzydale	lenie	tępaki
Azjaci	brzydale	rasiści	irytujący	głupcy	skąpcy

geje	źli	żywiący mylne poglądy	głupcy	irytujący	samoluby
chrześcijanie	głupcy	pomyleńcy	tępaki	cierpiący na urojenia	żywiący mylne poglądy

Widać w tym zestawieniu kilka stereotypów. Na przykład Afroamerykanie to jedyna grupa, do której przylgnęło miano „chamów". Prawie wszystkie pozostałe grupy etniczne to „głupcy", z wyjątkiem Żydów i muzułmanów; za ludzi złych z natury uważamy z kolei Żydów, muzułmanów i gejów, ale nie czarnych, Meksykanów, Azjatów czy chrześcijan.

Mianem terrorystów określa się wyłącznie muzułmanów. Kiedy ofiarą takiego stereotypu pada muzułmanin amerykański, reakcja społeczna może być natychmiastowa i okrutna. Baza danych wyszukiwań Google'a pozwala śledzić wybuchy napędzanej nienawiścią antyislamskiej wrogości minuta po minucie.

Zobaczmy, co się stało wkrótce po strzelaninie w San Bernardino w Kalifornii, 2 grudnia 2015 roku. Rankiem tego dnia Rizwan Farook i Tashfeen Malik przyszli na spotkanie kolegów Farooka z pracy uzbrojeni w półautomatyczne pistolety oraz karabiny i zamordowali czternaście osób. Wieczorem, dosłownie kilka minut po tym, jak media po raz pierwszy podały muzułmańsko brzmiące nazwisko jednego z napastników, niepokojąco znaczna liczba Kalifornijczyków już wiedziała, co należy zrobić z muzułmanami: pozabijać ich.

Najczęściej wpisywanym hasłem ze słowem „muzułmanin" było w tamtym czasie w Kalifornii „śmierć muzułmanom"[23]. A w całym kraju Amerykanie szukali frazy „śmierć muzułmanom" równie często jak fraz „przepis na martini", „objawy migreny" czy „skład Dallas Cowboys". Przez kilka dni po zamachu w San Bernardino na każdego Amerykanina bojącego się „islamofobii" przypadał inny Amerykanin szukający hasła „śmierć muzułmanom". Przed atakiem internauci

motywowani nienawiścią do wyznawców islamu stanowili w sieci dwadzieścia procent – po San Bernardino było ich ponad pięćdziesiąt.

Śledzenie przesyłu danych analitycznych minuta po minucie pokazuje, jak trudno jest uspokoić tego rodzaju wzburzenie. Cztery dni po strzelaninie ówczesny prezydent, Barack Obama, wygłosił telewizyjne orędzie do narodu. Z jednej strony chciał zapewnić Amerykanów, że rząd jest w stanie powstrzymać terrorystów, a z drugiej, co może było nawet ważniejsze, uciszyć niebezpieczną islamofobię.

Odwoływał się do lepszych cech natury ludzkiej, mówił o tym, jak ważne są tolerancja i poczucie wspólnoty. Jego retoryka była zarówno mocna, jak i wzruszająca. Dziennik „Los Angeles Times" chwalił Obamę za to, że „przestrzegał nas, aby strach nie zmącił nam rozumu". „New York Times" nazwał przemówienie „twardym", ale jednocześnie „uspokajającym". Portal Think Progress pochwalił je jako „niezbędny instrument dobrych rządów, zastosowany w celu ocalenia życia Amerykanom wyznania muzułmańskiego". Innymi słowy, przemówienie Obamy uznano za duży sukces. Czy było nim rzeczywiście?

Dane z Google'a zdają się przeczyć tej tezie. Przyjrzałem się im razem z Evanem Soltasem, który pracował wtedy na Uniwersytecie Princeton. Prezydent powiedział między innymi: „Wszyscy obywatele naszego kraju – bez względu na wyznanie – mają obowiązek odrzucić dyskryminację". Ale liczba wyszukiwań, w których nazywano muzułmanów „terrorystami", pisano, że są „niegodziwi", „skłonni do przemocy" i „źli", podwoiła się w czasie przemówienia i zaraz po nim. Obama powiedział też: „Mamy ponadto obowiązek odrzucić kryterium religijne przy podejmowaniu decyzji o tym, kogo wpuszczamy do naszego kraju". Mimo to liczba pejoratywnych haseł dotyczących syryjskich uchodźców, w przeważającej mierze muzułmanów, którzy rozpaczliwie szukali wtedy bezpiecznego schronienia, wzrosła o sześćdziesiąt procent, natomiast liczba wyszukiwań na temat tego, jak im pomóc, spadła o trzydzieści pięć. Obama prosił dalej Amerykanów, by „nie zapominali, że wolność jest silniejsza niż strach". Mimo to w trakcie jego przemówienia liczba wpisów w rodzaju „śmierć

muzułmanom" wzrosła trzykrotnie. Więcej, w czasie tego przemówienia i natychmiast po nim zwiększyła się liczba właściwie wszystkich typów obraźliwych wpisów na temat muzułmanów, jakie przyszło nam do głowy zadać, a liczba haseł pozytywnych spadła.

Wydawałoby się, że Obama zachował się tak, jak trzeba – wszystkie tradycyjne media gratulowały mu kojących słów. Ale nowe dane z Internetu, oferującego cyfrowe serum prawdy, sugerują, że przemówienie w rzeczywistości odniosło skutek przeciwny do zamierzonego. Zamiast, jak się wszystkim wydawało, uspokoić rozwścieczony motłoch, Obama wzburzył go tylko jeszcze bardziej, czego dowodzi analiza danych z sieci. Środki, które wydają się skuteczne, mogą przynosić skutki odwrotne do tych, na których nam zależy. I czasem potrzebujemy danych internetowych, by nie ulegać mimowolnej pokusie poklepywania samych siebie po plecach.

Co w takim razie powinien był powiedzieć prezydent, żeby ugasić tego rodzaju nienawiść, tak zaraźliwą dzisiaj w Ameryce? Wrócimy do tego później, teraz natomiast zajmiemy się najstarszą formą uprzedzenia w Stanach Zjednoczonych, czyli taką postacią nienawiści, która wyraźnie wyróżnia się spośród innych, jest chyba najbardziej destrukcyjna i od której zaczęliśmy tę książkę – bo w mojej pracy nad informacjami z bazy danych Google'a najbardziej bijącym w oczy przejawem hejtu wydało mi się niezwykle często używane słowo „czarnuch".

W liczbie pojedynczej albo mnogiej „czarnuch" co roku pojawia się w siedmiu milionach wyszukiwań internetowych w USA (i znów, w piosenkach raperskich prawie zawsze określenie to jest pisane jako „nigga", nie „nigger", stąd tych jego muzycznych zastosowań nie bierze się właściwie pod uwagę). Wyszukiwania „dowcipów o czarnuchach" zdarzają się siedemnaście razy częściej niż „dowcipów o żydkach", „żółtkach", „brudnych Latynosach", „chinolach" i „pedałach" razem wziętych[24].

Kiedy wyszukuje się najwięcej „dowcipów o czarnuchach"? Wtedy, gdy Afroamerykanie występują w programach informacyjnych. Wyjątkowo dużo takich wpisów pojawiło się na przykład po przejściu huraganu Katrina, gdy w telewizji i gazetach zaroiło się od materiałów

filmowych i zdjęć walczących o przeżycie, zrozpaczonych czarnych mieszkańców Nowego Orleanu. Liczba tego typu wyszukiwań wzrosła również w czasie pierwszej elekcji Obamy i zwiększa się zawsze średnio o trzydzieści procent w dniu upamiętniającym śmierć Martina Luthera Kinga[25].

Przerażająca wszechobecność takich etnicznych obelg podaje w wątpliwość niektóre nasze definicje rasizmu.

Każda teoria tego zjawiska musi wyjaśnić wielką zagadkę amerykańskiej historii. Z jednej strony przeważająca większość czarnoskórych Amerykanów uważa, że pada ofiarą uprzedzeń – i dysponuje wieloma dowodami na dyskryminację, która dotyka ich na komendach policji, w czasie rozmów o pracę czy na salach sądowych. Z drugiej strony bardzo niewielu białych Amerykanów przyznaje się do niechęci wobec czarnych.

W ostatnich czasach politolodzy są na ogół zgodni, że za ten stan rzeczy odpowiadają w dużej mierze szeroko rozpowszechnione uprzedzenia utajone. Nowa teoria naukowa głosi, że biali może i mają dobre intencje, ale są źle nastawieni do czarnych podświadomie, co wpływa na to, jak ich traktują. Uczeni wymyślili nawet pomysłowe badanie na obecność tego rodzaju nietolerancji – jest to tak zwany test utajonych skojarzeń.

Badania te niezmiennie potwierdzają, że większość z nas potrzebuje kilka milisekund więcej, żeby skojarzyć czarnoskóre twarze z pozytywnymi określeniami, takimi jak „dobry", niż z pejoratywnymi, takimi jak „okropny". Z twarzami białych jest natomiast odwrotnie. Ten dodatkowy czas potrzebny do zastanowienia to właśnie świadectwo utajonych uprzedzeń badanego – uprzedzeń, z których nawet on sam może nie zdawać sobie sprawy.

Istnieje jednak alternatywne wyjaśnienie tej dyskryminacji, której Afroamerykanie doznają, a biali zaprzeczają: skrywany jawny rasizm. Przypuśćmy, że istnieje dość szeroko rozpowszechniony rasizm uświadomiony, z którego ludzie doskonale zdają sobie sprawę, ale do którego nie chcą się przyznać, a już na pewno nie w żadnym sondażu. Dane analityczne wskazują, że tak właśnie jest. W wyszukiwaniu hasła „dowcipy o czarnuchach" nie ma nic utajonego. Trudno także sobie

wyobrazić, by biali Amerykanie, wpisując w wyszukiwarkę słowo „czarnuch" równie często jak „migrena" i „ekonomista", nie przejawiali w ten sposób jawnego rasizmu i nie oddziaływali w poważny sposób na postawę Afroamerykanów. Zanim dostaliśmy do ręki dane Google'a, nie mieliśmy sposobu, by jakoś zmierzyć tę zaraźliwą niechęć. Ale skoro teraz już go mamy, możemy się przekonać, co ona nam wyjaśnia.

Wyjaśnia zaś na przykład, jak powiedzieliśmy już wcześniej, dlaczego wyniki wyborcze prezydenta Obamy w wielu regionach kraju były niższe od prognozowanych. Istnieje tutaj także korelacja z różnicą w płacach między białymi i czarnymi obywatelami Stanów Zjednoczonych[26], o której niedawno pisał zespół ekonomistów, okazało się bowiem, że tam, gdzie pojawia się najwięcej rasistowskich wyszukiwań, czarni otrzymują zaniżone wynagrodzenie. No i mamy jeszcze kandydaturę Trumpa. We wstępie wspominałem, że kiedy Nate Silver, guru sondaży, szukał zmiennej geograficznej, którą można by dopasować do poparcia, jakie Trump uzyskał w prawyborach, znalazł ją na sporządzonej przeze mnie mapie rasizmu – zmienną tą okazały się bowiem wyszukiwania internetowe zawierające określenie „czarnuch".

Uczeni przeprowadzili niedawno pomiary skali utajonych uprzedzeń wobec czarnych – dla każdego stanu osobno – dzięki czemu mogłem porównać skutki jawnego rasizmu, mierzonego według wyszukiwań w Google'u, i skutki uprzedzeń skrywanych. Sprawdziłem na przykład, jak silnie jedne i drugie działały przeciwko Obamie w czasie obu wygranych przez niego elekcji. Zastosowałem analizę regresywną i okazało się, że dzięki badaniom rasistowskich wyszukiwań w sieci można było w większości przypadków przewidzieć, gdzie Barack Obama osiągnie wynik słabszy od przewidywanego – natomiast testy utajonych skojarzeń tłumaczyły w tej kwestii bardzo niewiele.

Dyskryminację, której regularnie doświadczają czarni w Ameryce, wydaje się zatem napędzać raczej uświadomiona, chociaż ukrywana niechęć do nich, lecz w innych grupach społecznych uprzedzenia podświadome mogą odgrywać bardziej znaczącą rolę. Udało mi się między innymi wykorzystać wpisy w Google'u, aby znaleźć dowody na

istnienie utajonych uprzedzeń wobec innego segmentu amerykańskiej populacji: młodych dziewcząt.

A kto, możecie zapytać, miałby żywić wobec nich uprzedzenia?

Rodzice[27].

Trudno się dziwić, że matki i ojcowie często reagują emocjonalnie na myśl, że mogą mieć utalentowane dzieci. Prawie wszystkie wyszukiwania w Google'u zaczynające się od „Czy moje dwuletnie dziecko…" kończy fraza „jest utalentowane". Pytanie to nie bywa jednak zadawane równie często w odniesieniu do chłopców i dziewczynek. Rodzice dwa i pół razy częściej pytają „Czy mój syn ma talent?" niż „Czy moja córka ma talent?". Wykazują też podobne uprzedzenia, zadając inne pytania związane z inteligencją dziecka, które wstydziliby się może sformułować na głos, jak na przykład „Czy mój syn jest geniuszem?".

Czy dostrzegają jakieś prawdziwe różnice między dziewczynkami i chłopcami? Może chłopcy częściej używają trudnych słów albo jakoś inaczej, chociaż obiektywnie, zdradzają takie czy inne zdolności? Nie. Jeżeli naprawdę istnieje jakaś różnica, to raczej na korzyść dziewczynek, które, co pokazują badania, w młodym wieku prawie zawsze dysponują większym słownictwem niż chłopcy i używają bardziej złożonych zdań. W szkołach amerykańskich dziewczynki mają dziewięć procent więcej szans na to, żeby trafić do klas dla szczególnie uzdolnionych uczniów[28]. Mimo to rodzice przy stole w czasie obiadu częściej dostrzegają utalentowanych chłopców niż dziewczęta[12*]. We wszystkich badanych przeze mnie wpisach odnoszących się do inteligencji pytali raczej o chłopców, chociaż częściej zadają też pytania w rodzaju „Czy mój syn nie daje sobie rady?" albo „Czy mój syn jest głupi?". Ale wpisy z pejoratywnymi określeniami w rodzaju „nie daje sobie rady" lub „jest głupi" nie są aż tak chronicznie ograniczone do synów, jak wpisy z uwagami pozytywnymi typu „jest utalentowany" czy też „jest geniuszem".

Czym zatem najbardziej przejmują się rodzice w odniesieniu do córek? Przede wszystkim ich wyglądem zewnętrznym. Weźmy wpisy na temat wagi dziecka. Pytania w rodzaju „Czy moja córka ma nadwagę?"

pojawiają się mniej więcej dwa razy częściej niż takie same pytania o synów, podobnie jak pytania o to, w jaki sposób skłonić córki, żeby się odchudzały. Ale tak jak w wypadku talentu, tego rodzaju uprzedzenia nie znajdują odzwierciedlenia w rzeczywistości – na nadwagę cierpi bowiem tylko dwadzieścia osiem procent dziewcząt w Ameryce, a chłopców trzydzieści pięć procent[29]. Tymczasem chociaż pomiary mówią nam wyraźnie, że mamy w kraju więcej otyłych uczniów niż uczennic, rodzice znacznie częściej widzą dziewczynki z nadwagą niż chłopców – lub może po prostu bardziej się o nie martwią.

Rodzice półtora raza częściej szukają potwierdzenia, że ich córka jest piękna – rzadziej, że mają przystojnego syna, a o ewentualną brzydotę dziewczyny pytają w sieci trzy razy częściej niż o brzydotę chłopca (nie mam pojęcia, skąd Internet miałby wiedzieć, czy dziecko jest brzydkie, czy ładne).

Generalnie rzecz biorąc, rodzice częściej używają więc pozytywnych określeń, gdy pytają o synów. Są przy tym także bardziej skłonni pytać o to, czy ich syn jest „szczęśliwy", niż czy „jest w depresji".

Liberalni czytelnicy mogliby sądzić, że takie uprzedzenia występują częściej w konserwatywnych rejonach kraju, ale nie znalazłem żadnych dowodów, które by potwierdzały taką tezę. Prawdę mówiąc, nie znalazłem żadnego ważniejszego związku między którymkolwiek z powyższych uprzedzeń i politycznym albo kulturowym wizerunkiem danego stanu. Brakuje też dowodów, że częstotliwość występowania uprzedzeń płciowych wśród rodziców zmniejszyła się od 2004 roku, w którym po raz pierwszy udostępniono bazę danych Google'a. Wygląda zatem na to, że uprzedzenie wobec dziewcząt jest bardziej rozpowszechnione i zakorzenione głębiej, niż moglibyśmy się tego spodziewać.

Seksizm nie jest jedyną sferą naszego życia, w której stereotypy mogą okazać się mylące.

VikingMaiden88 ma dwadzieścia sześć lat. Lubi czytać książki historyczne i pisze wiersze. Jej ulubiony cytat, którym przyozdabia swoje wpisy na forach, pochodzi z Shakespeare'a. Dowiedziałem się

tego z jej profilu i postów na stormfront.org, najpopularniejszym w Ameryce portalu hejterskim. Dowiedziałem się również, że VikingMaiden88 spodobał się pewien artykuł na portalu gazety, dla której pracuję – „New York Timesa". Wyrażała się o nim wręcz entuzjastycznie.

W ostatnim czasie przeanalizowałem dziesiątki tysięcy profili na Stormfroncie[30], którego zarejestrowani członkowie mogą podać swoje miejsce zamieszkania, datę urodzenia, zainteresowania oraz inne dane.

Stormfront został założony w 1995 roku przez Dona Blacka, byłego przywódcę Ku-Klux-Klanu. Najpopularniejsze „grupy społecznościowe" na tym portalu to „Związek Narodowych Socjalistów" i „Fani oraz zwolennicy Adolfa Hitlera". Jak podaje serwis Quantcast, w ubiegłym roku każdego miesiąca portal Blacka odwiedzało od dwustu do czterystu tysięcy Amerykanów. Z niedawnego raportu Southern Poverty Law Center wynika, że w ostatnich pięciu latach zarejestrowani członkowie Stormfrontu zamieszani byli w blisko sto morderstw.

Nigdy nie domyśliłbym się tożsamości tych ludzi.

To przede wszystkim młodzież, w każdym razie jeśli wierzyć podawanym w sieci datom urodzenia. Na portal najczęściej zapisują się dziewiętnastolatkowie, których jest wśród członków czterokrotnie więcej niż czterdziestolatków. Owszem, internauci i użytkownicy portali społecznościowych w ogóle są zwykle młodzi, ale nie aż tak.

W profilach członkowskich nie wymaga się podania płci. Przejrzałem jednak wszystkie posty i biogramy wybranej losowo próbki użytkowników portalu i okazało się, że można na tej podstawie ustalić płeć większości z nich – szacuję więc, że około trzydziestu procent członków Stormfrontu to kobiety.

Najwięcej pochodzi z Montany, Alaski i Idaho. Są to stany zamieszkane w przytłaczającej większości przez białych. Czy to oznacza, że dorastanie w otoczeniu mało zróżnicowanym etnicznie rodzi nienawiść?

Chyba nie. Jest raczej tak, że ponieważ te stany zamieszkuje proporcjonalnie większa liczba białych nie-Żydów, to właśnie taka ludność stanowi potencjalne zaplecze grupy, atakującej Żydów

i niebiałych. Procentowo spośród użytkowników Stormfrontu, którzy z czasem zostają jego członkami, większość pochodzi z takich rejonów kraju, gdzie żyje więcej mniejszościowych grup etnicznych. Dotyczy to zwłaszcza tych aktywistów organizacji, którzy mają osiemnaście lat lub mniej, zwykle nie mogą więc jeszcze samodzielnie decydować o tym, gdzie mieszkają.

Na przykład członków z Kalifornii – gdzie odsetek mniejszości etnicznych należy do największych w całym kraju – znajdziemy w tej grupie wiekowej o dwadzieścia pięć procent więcej, niż wynosi średnia krajowa Stormfrontu.

Jedną z najpopularniejszych grup społecznościowych na portalu tworzą, jak sami siebie określają, „zwolennicy antysemityzmu". Odsetek członków organizacji Blacka, którzy wstępują właśnie do tej grupy, jest ściśle skorelowany z liczebnością populacji żydowskiej w danym stanie. Najwięcej stormfrontowców pochodzi ze stanu Nowy Jork, gdzie mieszka najwięcej Żydów w USA.

W 2001 roku do Stormfrontu wstąpił Dna88, przedstawiający się jako „przystojny i społecznie świadomy" trzydziestoletni twórca stron internetowych, zamieszkały w „Żydowym Jorku". W ciągu następnych czterech miesięcy napisał ponad dwieście postów, takich jak „Żydowskie zbrodnie przeciwko ludzkości" i „Splamiona krwią żydowska forsa", przekierowując także użytkowników na stronę jewwatch.com, która podaje się za „naukową bibliotekę", gromadzącą teksty poświęcone „przestępczości syjonistycznej".

Członkowie Stormfrontu narzekają, że mniejszości etniczne nie znają angielskiego i popełniają przestępstwa. Najbardziej jednak zaciekawiły mnie ich skargi na konkurencję na rynku… randkowym.

Pewien facet, używający nazwiska William Lyon Mackenzie King – na cześć byłego premiera Kanady, który stwierdził kiedyś, że „Kanada powinna pozostać krajem białych ludzi" – napisał w roku 2003, że musiał „hamować gniew" na widok białej kobiety „noszącej na rękach swoje brzydkie, skundlone, półmurzyńskie świniątko". Z kolei Whitepride26, czterdziestojednoletnia studentka z Los Angeles, stwierdza na swoim profilu: „Nie lubię czarnych, Latynosek i czasami

Azjatek, zwłaszcza kiedy faceci uważają je za atrakcyjniejsze od białych kobiet".

Swoją rolę w tej sferze odgrywają również wydarzenia polityczne. Zdecydowanie najwięcej nowych członków zapisało się do Stormfrontu 5 listopada 2008 roku, dzień po wyborze Baracka Obamy na prezydenta Stanów Zjednoczonych. W czasie kandydatury Donalda Trumpa[31] nie odnotowano natomiast zwiększonego zainteresowania Stormfrontem, a liczba nowych zapisów wzrosła nieznacznie na chwilę tylko bezpośrednio po jego zwycięstwie. Trumpa wyniosła na szczyt fala białego nacjonalizmu, brakuje jednak dowodów, że to on ją wywołał.

Biały nacjonalizm przybrał na sile po wygranej Obamy. A zwycięstwo Trumpa było tego efektem.

Nieistotna w tej kwestii wydaje się natomiast gospodarka. Nie widać żadnego związku między miesięczną liczbą nowych rejestracji na portalu Blacka i stopą bezrobocia. W stanach najboleśniej dotkniętych wielką recesją liczba wyszukiwań stron Stormfrontu właściwie się nie zwiększyła.

Najbardziej interesujące, a może i zaskakujące, są jednak niektóre tematy poruszane przez członków tej organizacji na forach dyskusyjnych, bo sam rozmawiam o podobnych sprawach ze swoimi znajomymi. Być może jestem naiwny, ale wyobrażałem sobie raczej, że biali nacjonaliści żyją w innym świecie niż ja i moi przyjaciele. Tymczasem stormfrontowcy prowadzą długie dyskusje, wychwalające *Grę o tron*, i porównują zalety internetowych portali randkowych, takich jak PlentyOfFish czy OkCupid.

A oto najważniejszy fakt dowodzący tego, że użytkownicy Stormfrontu żyją mimo wszystko w świecie podobnym do świata mojego i moich najbliższych: otóż w ich środowisku dużą popularnością cieszy się „New York Times". Nie tylko VikingMaiden88 odwiedza naszą stronę – moją gazetę czyta wielu białych nacjonalistów. W porównaniu z czytelnikami Yahoo News członkowie Stormfrontu odwiedzają stronę „NYT" dwa razy chętniej od nich.

Więc użytkownicy hejterskiego portalu przeglądają ten obrzydliwie liberalny nytimes.com? Jak to możliwe? Jeżeli tak znaczna liczba

stormfrontowców czytuje wiadomości z „NYT”, to mylimy się w naszej ocenie białych nacjonalistów. Mylimy się też co do tego, jak naprawdę funkcjonuje Internet w naszym społeczeństwie.

PRAWDA O INTERNECIE

Chyba niemal wszyscy zgodzimy się, że Internet w pewnym sensie dzieli Amerykanów – z reguły zaszywamy się bowiem wszyscy na portalach, których użytkownikami są osoby wyznające poglądy zbieżne z naszymi. Cass Sunstein z Harvard Law School opisuje tę sytuację tak: „Rynek komunikacji zmierza raptownie tam, gdzie ludzie ograniczają się do własnych opinii – liberałowie obserwują i czytają w większości albo wyłącznie innych liberałów, osoby o umiarkowanych poglądach czytają posty osób o umiarkowanych poglądach, konserwatyści czytają konserwatystów, a neonaziści neonazistów”.

Ta teza wydaje się słuszna. Internet zapewnia nam przecież właściwie nieograniczoną liczbę źródeł, z których możemy czerpać wiadomości o świecie. I ja, i wy możemy czytać, na co tylko mamy ochotę, podobnie jak VikingMaiden88, choć ludzie, jeśli nie wywierać na nich żadnego nacisku, szukają na ogół takich zapatrywań, które potwierdzają to, w co sami wierzą. Jeżeli tak jest naprawdę, to Internet z pewnością prowadzi do radykalnych podziałów politycznych[32].

Mam tylko jeden problem z tym potocznym przeświadczeniem: otóż analiza danych mówi nam po prostu, że nie jest ono prawdziwe.

Dowodów zaprzeczających, że owa powszechnie obowiązująca mądrość zgadza się z faktami, dostarczają choćby pochodzące z 2011 roku badania Matta Gentzkowa i Jessego Shapiro, ekonomistów, o których pracach mówiliśmy już wcześniej.

Gentzkow i Shapiro sprawdzili najpierw, czego reprezentatywna grupa Amerykanów szukała w sieci pod kątem deklarowanych przez nich liberalnych albo konserwatywnych poglądów, a zgromadziwszy dane na ten temat, spróbowali zmierzyć polityczne podziały w Internecie.

W jaki sposób? Przeprowadzili interesujący eksperyment myślowy.

Załóżmy, że wybieramy losowo dwóch Amerykanów, którzy odwiedzają ten sam portal informacyjny. Jakie będzie prawdopodobieństwo, że jeden z nich okaże się liberałem, a drugi konserwatystą? Inaczej mówiąc, jak często liberałowie i konserwatyści „spotykają się" na takich portalach?

I dalej – przypuśćmy, że liberałowie i konserwatyści nigdy nie czerpią wiadomości w Internecie z tego samego źródła, czyli że liberałowie odwiedzają wyłącznie portale liberalne, a konserwatyści tylko konserwatywne. Gdyby faktycznie tak było, prawdopodobieństwo, że dwóch Amerykanów odwiedzających ten sam portal informacyjny wyznaje sprzeczne poglądy polityczne, wynosiłoby zero. W takiej sytuacji Internet podzielony byłby doskonale, a liberałowie i konserwatyści nigdy nie spotykaliby się ze sobą w sieci.

Przypuśćmy też dla kontrastu, że liberałowie i konserwatyści nie różnią się ani trochę, jeżeli chodzi o miejsca zdobywania informacji w sieci, czyli że z równym prawdopodobieństwem mogą odwiedzić dowolny portal dziennikarski. Gdyby sytuacja tak wyglądała, to prawdopodobieństwo, że dwaj Amerykanie odwiedzający jeden portal mają sprzeczne poglądy polityczne, wynosiłoby mniej więcej pięćdziesiąt procent. Wtedy Internet okazałby się doskonale homogeniczny, a liberałowie i konserwatyści idealnie mieszaliby się ze sobą w sieci.

Co zatem mówi nam w tej sprawie analiza danych? Zdaniem Gentzkowa i Shapiro w Stanach Zjednoczonych prawdopodobieństwo, że dwie osoby odwiedzające ten sam portal dziennikarski wyznają różne poglądy polityczne, wynosi mniej więcej czterdzieści pięć procent, co oznacza, że Internet jest znacznie bliższy doskonałej jednorodności niż doskonałemu podziałowi, a liberałowie i konserwatyści „spotykają się" ze sobą w sieci przez cały czas.

Możemy lepiej zrozumieć, co oznacza brak podziału w Internecie, jeśli przyjrzymy się podziałom w pozostałych sferach naszego codziennego życia. Gentzkow i Shapiro zdołali powtórzyć swoją analizę w kontekście innego rodzaju interakcji offline. Jakie istnieje

prawdopodobieństwo, że sprzeczne poglądy polityczne żywią dwaj członkowie jednej rodziny? Sąsiedzi? Koledzy z pracy? Znajomi?

Korzystając z danych z General Social Survey, nasi uczeni stwierdzili, że we wszystkich wymienionych przypadkach prawdopodobieństwo, iż badani wyznają sprzeczne poglądy polityczne, jest mniejsze niż w Internecie.

Prawdopodobieństwo, że spotkana przez ciebie osoba ma inne poglądy polityczne niż ty

Na portalu informacyjnym	45,2%
W miejscu pracy	41,6%
Wśród sąsiadów	40,3%
Wśród członków rodziny	37,0%
Wśród znajomych	34,7%

Wynika z tego, że jest bardziej prawdopodobne, że spotkasz kogoś o poglądach politycznych sprzecznych z twoimi w Internecie niż offline.

Dlaczego Internet nie jest bardziej podzielony? Istnieją dwa czynniki, które ograniczają polityczne podziały w sieci.

Po pierwsze, co może nieco dziwić, internetowy przemysł informacyjny został zdominowany przez kilka potężnych portali. Myślimy zazwyczaj, że sieć przyciąga skrajności – i owszem, znajdziecie w niej portale dla każdego bez względu na poglądy. Istnieją strony dla zwolenników powszechnego dostępu do broni i jego przeciwników, dla amatorów cygar i aktywistów na rzecz wprowadzenia do obiegu monet jednodolarowych, dla anarchistów i białych nacjonalistów. Lecz wszystkie takie portale stanowią zaledwie drobinkę w świecie wiadomości internetowych – bo w roku 2009 ponad połowę informacji w sieci przekazywały jedynie cztery portale dziennikarskie: Yahoo News, AOL News, msnbc.com i cnn.com. Najpopularniejszy wśród

Amerykanów pozostaje Yahoo News, który odwiedza blisko dziewięćdziesiąt milionów gości miesięcznie, czyli blisko sześćset razy więcej niż Stormfront. Portale medialne takie jak Yahoo News przyciągają szeroką i politycznie zróżnicowaną publiczność.

Drugi powód, dla którego Internet nie jest aż tak podzielony, jak mogłoby się wydawać, wynika stąd, że wiele osób o jasno określonych poglądach politycznych odwiedza portale reprezentujące inne punkty widzenia – choćby po to, żeby się nakręcić i podjąć z nimi polemikę. Polityczni nałogowcy nie ograniczają się wyłącznie do portali, które celują w nich jako w swoich odbiorców. Ktoś, kto odwiedza thinkprogress.org czy moveon.org, portale skrajnie liberalne, wchodzi na zdradzający odchylenie prawicowe foxnews.com częściej niż przeciętny użytkownik Internetu; a ten, kto odwiedza rushlimbaugh.com i glennback.com, portale skrajnie konserwatywne, częściej niż przeciętny internauta odwiedza też nytimes.com, witrynę zdecydowanie bardziej liberalną.

Badania Gentzkowa i Shapiro opierały się na danych z lat 2004–2009, czyli ze stosunkowo wczesnego okresu istnienia Internetu – który od tamtej pory może się bardziej podzielił? Czyż nie zmieniły go media społecznościowe, zwłaszcza Facebook? Przecież jeśli nasi znajomi wyznają takie same przekonania polityczne jak my, to media społecznościowe powinny je odzwierciedlać i powtarzać niczym echo, prawda?

I tym razem jednak sprawa okazuje się nie taka prosta. Chociaż prawdą jest, że nasi znajomi na Facebooku raczej podzielają nasze opinie, to zespół analityków – Eytan Bakshy, Solomon Messing i Lada Adamic – dowiódł, że zaskakująco wiele informacji, które czerpiemy z Facebooka, pochodzi od osób o poglądach przeciwstawnych do naszych[33].

Jak to możliwe? Przecież znajomi chyba dzielą z nami przekonania ideologiczne? Owszem, tak, jest wszakże jeden bardzo istotny powód, dla którego Facebook może prowadzić do bardziej zróżnicowanej dyskusji politycznej niż spotkania towarzyskie offline[34]. Otóż przeciętnie mamy o wiele więcej znajomych na Facebooku niż

w realnym świecie[35] – a wśród tych słabo nam znanych użytkowników Facebooka częściej spotkamy ludzi, którzy nie zgadzają się z naszymi zapatrywaniami.

Facebook tworzy słabe więzi społeczne[36] – bo utrzymujesz na nim kontakty z jakimś znajomym z liceum, stukniętą daleką kuzynką albo kumplem kumpla twojego kumpla, którego, no może rzeczywiście kiedyś raz w życiu gdzieś widziałeś. Ale z tymi ludźmi chyba raczej nigdy nie pójdziesz pograć w kręgle ani na grilla. Nie zaprosisz ich na kolację, chociaż są twoimi znajomymi na Facebooku, gdzie widzisz podane przez nich linki do artykułów, zawierających opinie, którymi inaczej może nigdy nie zawracałbyś sobie głowy.

Podsumowując, możemy stwierdzić, że Internet w gruncie rzeczy zbliża ludzi o różnych przekonaniach. Przeciętny liberał spędza zapewne poranki ze swoją liberalną żoną i liberalnymi dziećmi, popołudnia z liberalnymi kolegami w pracy, do której jeździ w otoczeniu samochodów z liberalnymi naklejkami na zderzakach, a wieczorami ćwiczy jogę w otoczeniu innych liberalnych kursantów – więc kiedy po powrocie do domu przegląda kilka konserwatywnych komentarzy na cnn.com albo dostaje facebookowego linka od jakiegoś znajomego z czasów szkolnych, który hołduje republikańskim poglądom, to może być tego dnia jego jedyny kontakt z konserwatywnymi przekonaniami politycznymi.

W mojej ulubionej kawiarni w Brooklynie raczej nie spotykam białych nacjonalistów, ale zarówno VikingMaiden88, jak i ja chętnie odwiedzamy portal „New York Timesa”.

PRAWDA O PRZEMOCY WOBEC DZIECI I ABORCJI

Internet może nam ujawnić nie tylko niepokojące ludzkie postawy, ale też zachowania. Więcej – dane z Google’a mogą skutecznie zwracać naszą uwagę na pewne sytuacje kryzysowe, które tradycyjne źródła informacji zwykle pomijają milczeniem. A ludzie przecież zwracają się do Google’a, gdy mają jakieś problemy.

Weźmy pod uwagę molestowanie dzieci w czasie wielkiej recesji.

Kiedy pod koniec 2007 roku zaczął się poważny kryzys gospodarczy, wielu ekspertów z oczywistych powodów martwiło się tym, w jaki sposób odbije się on na dzieciach, bo przecież niejeden rodzic mógł szybko popaść z jego powodu w stres i depresję, a te czynniki zwiększają ryzyko molestowania dzieci oraz stosowania siły wobec nich. Dlatego przypuszczano, że przemoc wobec osób małoletnich może się znacznie nasilić.

Potem napłynęły jednak oficjalne dane i wydawało się, że obawy specjalistów były nieuzasadnione. Organizacje opiekuńcze donosiły, że mają mniej zgłoszeń przypadków przemocy i molestowania, a ich liczba spadła najbardziej w stanach najsilniej dotkniętych recesją. „Złowieszcze i złowróżbne prognozy nie ziściły się”[37], stwierdził w 2011 roku Richard Gelles, ekspert z Uniwersytetu Pensylwanii, w rozmowie z przedstawicielami Associated Press. Fakty, w każdym razie oficjalnie, nie potwierdzały bowiem zdroworozsądkowych domysłów; przemoc wobec dzieci w czasie recesji spadła.

Ale czy rzeczywiście, skoro tylu dorosłych pozostawało bez pracy i żyło w skrajnym napięciu nerwowym? Trudno mi było w to uwierzyć, postanowiłem więc przyjrzeć się danym z Google’a.

Okazało się, że wiele dzieci zamieszcza w wyszukiwarce dramatyczne i zatrważające wpisy, takie jak „mama mnie bije” albo „tata mnie uderzył”[38]. Wyszukiwania sieciowe przedstawiają inny, niepokojący obraz wydarzeń z tamtych czasów – liczba tego rodzaju wpisów znacznie się podniosła w latach recesji i była ściśle powiązana z wysokością stopy bezrobocia.

Powiem wam, co stało się moim zdaniem: otóż spadła jedynie liczba doniesień o przypadkach stosowania przemocy wobec dzieci, nie liczba takich przypadków. Zresztą ocenia się, że i tak zgłaszany jest zaledwie ich skromny odsetek, a wiele osób, które dokonują zwykle podobnych zgłoszeń, jak nauczyciele czy policjanci (albo je rozpatrują, jak pracownicy rozmaitych agencji opiekuńczych), jest w okresie recesji przepracowanych lub pozostaje bez pracy.

Mamy zresztą więcej dowodów na to, tym razem nie z Google'a, że w czasie kryzysu przemoc wobec osób małoletnich w rzeczywistości wzrosła. Jeżeli dziecko umrze na skutek przemocy albo zaniedbania, zgon musi zostać zgłoszony, a liczba takich śmiertelnych przypadków, chociaż niewielka, podniosła się jednak właśnie w stanach najpoważniej dotkniętych recesją.

Istnieją również kolejne dowody, pochodzące z Google'a, że w tych rejonach kraju więcej osób niż zwykle podejrzewało wzrost liczby przypadków stosowania siły wobec dzieci. Biorąc pod uwagę wyniki sprzed kryzysu i trend ogólnokrajowy, stwierdziłem, że w stanach, które ucierpiały relatywnie najbardziej, zwiększyła się liczba wpisów dotyczących zaniedbań wychowawczych i przemocy wobec osób małoletnich. Na każdy kolejny punkt procentowy wzrostu stopy bezrobocia liczba wyszukiwań o treści „przemoc wobec dzieci" lub „zaniedbane dziecko" wzrastała o trzy procent. Należy jednak przypuszczać, że większość internautów dokonujących takich wpisów nigdy skutecznie nie zgłosiła żadnego aktu przemocy, bo właśnie w tych stanach liczba tego rodzaju zgłoszeń spadła.

Więc liczba wpisów ze strony cierpiących dzieci rośnie. Wzrasta liczba zgonów wśród najmłodszych. W najbardziej dotkniętych kryzysem stanach zwiększa się liczba wyszukiwań ze strony osób podejrzewających, że w ich otoczeniu dochodzi do aktów przemocy. Lecz liczba ich zgłoszeń spada. Wydaje się, że wskutek recesji coraz więcej dzieci zwierza się w Internecie, że są bite przez rodziców, a zarazem coraz więcej osób sądzi, że dostrzega przemoc w swoim najbliższym otoczeniu. Przepracowane agencje opiekuńcze rozpatrują jednak mniej takich spraw niż zwykle[39].

Uważam zatem, że można powiedzieć, iż kryzys gospodarczy zwiększył liczbę przypadków przemocy wobec dzieci, chociaż tradycyjne metody badawcze nie zarejestrowały tego zjawiska.

Ostatnio zawsze zwracam się do Google'a, jeżeli podejrzewam, że są gdzieś ludzie, którzy cierpią, powiedzmy, „nieoficjalnie". To jedna z potencjalnych korzyści płynących ze zbiorów nowych danych oraz

faktu, że umiemy je interpretować, otrzymujemy bowiem możliwość pomocy osobom słabszym i wrażliwym, o których inaczej władza mogłaby zapomnieć.

Dlatego gdy niedawno Sąd Najwyższy przypatrywał się skutkom nowego prawa, utrudniającego przeprowadzanie zabiegów aborcyjnych, zwróciłem się ku danym z wyszukiwarki, ponieważ przypuszczałem, że kobiety, których ów akt legislacyjny dotyczył, mogą szukać w sieci informacji o tym, jak nielegalnie dokonać przerwania ciąży[40]. I rzeczywiście, szukały. A liczba wpisów na ten temat była najwyższa w stanach, które przegłosowały prawo ograniczające możliwość aborcji.

Dane analityczne w tej kwestii są zarówno użyteczne, jak i niepokojące.

W 2015 roku w amerykańskim Google'u pojawiło się ponad siedemset tysięcy wyszukiwań metod dokonania samodzielnej aborcji, podczas gdy liczba wpisów ze strony osób poszukujących klinik aborcyjnych przekroczyła trzy miliony czterysta tysięcy. Wyraźnie świadczy to o tym, że spory odsetek kobiet, które rozważały przerwanie ciąży, brał pod uwagę możliwość dokonania zabiegu własnymi rękami.

Kobiety blisko sto sześćdziesiąt tysięcy razy szukały sposobów zdobycia pigułek aborcyjnych nieoficjalnymi kanałami, wpisując takie hasła, jak „kupię pigułki aborcyjne online" albo „pigułki aborcyjne za darmo". Pytały też Google'a o możliwość dokonania aborcji z zastosowaniem ziół, takich jak natka pietruszki, albo witaminy C. Cztery tysiące razy szukano instrukcji, jak przeprowadzić aborcję drucianym wieszakiem na ubrania, w tym tysiąc trzysta wpisów sformułowano dosłownie: „jak zrobić aborcję wieszakiem". Kilkaset pytań dotyczyło możliwości przeprowadzenia aborcji przez wprowadzenie do macicy wybielacza albo wskutek serii uderzeń w brzuch.

Co napędza zainteresowanie przeprowadzeniem aborcji na własną rękę? Dane geograficzne i czas wyszukiwań w Google'u wskazują prawdopodobnego winnego: gdy trudno jest dokonać aborcji zgodnie z prawem, kobiety szukają możliwości nielegalnych.

Pomiędzy 2004 i 2007 rokiem liczba wyszukiwań dotyczących domowych sposobów przerywania ciąży była stosunkowo stała. Zaczęła się podnosić dopiero pod koniec 2008 roku, czyli mniej więcej wtedy, kiedy doszło do kryzysu finansowego, a później gospodarczego, podskoczyła zaś najbardziej w roku 2011 – aż o czterdzieści procent. Guttmacher Institute, organizacja proaborcyjna, wskazuje na rok 2011 jako początek antyaborcyjnego zwrotu w USA – wtedy to bowiem w poszczególnych stanach wprowadzono dziewięćdziesiąt dwie różne regulacje ograniczające dostęp do aborcji. Jeśli dla porównania spojrzymy na Kanadę, gdzie podobny atak antyaborcyjny nie nastąpił, to okaże się, że w omawianym okresie nie odnotowano tam porównywalnego wzrostu zainteresowania domowymi metodami przerywania ciąży.

Najwięcej wyszukiwań aborcyjnych zanotowano w Missisipi, stanie zamieszkanym przez w przybliżeniu trzy miliony osób i mającym dzisiaj zaledwie jedną klinikę aborcyjną. Osiem z dziesięciu stanów o najwyższym wskaźniku wpisów dotyczących samodzielnie przeprowadzonego zabiegu przerywania ciąży Guttmacher Institute uważa za wrogie lub bardzo wrogie aborcji. Nie zalicza do nich jednak żadnego z dziesięciu stanów o najniższym wskaźniku takich wyszukiwań.

Oczywiście na podstawie wpisów w Google'u nie możemy wiedzieć, ile kobiet poddaje się własnoręcznie przeprowadzonej aborcji, dowody analityczne sugerują jednak, że ich liczba może być wysoka. Spróbujmy naświetlić ten problem, zestawiając ze sobą dane dotyczące aborcji i urodzin.

Porównajmy dziesięć stanów, w których znajduje się najwięcej klinik aborcyjnych w przeliczeniu na jednego mieszkańca (są na tej liście Nowy Jork i Kalifornia), z dziesięcioma stanami, gdzie klinik aborcyjnych jest relatywnie najmniej (tu mamy między innymi Missisipi i Oklahomę). Kobiety mieszkające w stanach, gdzie klinik aborcyjnych jest jak na lekarstwo, przeprowadziły o pięćdziesiąt cztery procent mniej legalnych aborcji – o jedenaście zabiegów na każdy tysiąc kobiet w wieku pomiędzy piętnastym i czterdziestym czwartym rokiem życia. Amerykanki mieszkające tam, gdzie klinik jest najmniej, urodziły

również najwięcej dzieci, chociaż nie tyle, żeby zrównoważyć liczbę aborcji – na każdy tysiąc kobiet w wieku rozrodczym na świat przyszło w takich stanach sześcioro dzieci.

Wygląda więc na to, że w tych rejonach kraju, gdzie trudno jest dokonać zabiegu aborcji, nie wszystkie ciąże zostały odnotowane. Oficjalne źródła nie mówią nam jednak, co się stało z pięcioma brakującymi porodami na każdy tysiąc mieszkanek stanów antyaborcyjnych.

Tylko Google daje nam co do tego niezłe wskazówki.

Nie powinniśmy ślepo ufać oficjalnym danym. Rząd może nam powiedzieć, że liczba aborcji albo przypadków przemocy wobec dzieci spadła, a politycy mogą ten sukces odtrąbić. Ale wyniki, które rzekomo widzimy, powstają czasem wskutek wad w metodach zbierania danych. Prawda może być inna – niekiedy o wiele mroczniejsza.

PRAWDA O WASZYCH ZNAJOMYCH Z FACEBOOKA

Moja książka traktuje generalnie o big data, ale ten rozdział eksponuje głównie wpisy w Google'a, które, jak twierdzę, odsłaniają przed nami ukryty świat, bardzo różny od tego, jaki pozornie widzimy. Czy w takim razie inne źródła ze zbioru big data także są cyfrowym serum prawdy? Nie, wiele z nich, na przykład Facebook, to jego przeciwieństwo.

W mediach społecznościowych, tak jak i sondażach, brakuje nam motywacji, aby mówić prawdę. Lecz w mediach społecznościowych o wiele częściej niż w sondażach mamy poważne powody, żeby przedstawiać się innym z jak najlepszej strony. Ostatecznie twoja obecność online nie jest anonimowa. W sieci wdzięczysz się przecież do publiczności i mówisz swoim przyjaciołom, znajomym, krewnym, kolegom z pracy i nieznajomym, kim jesteś.

Żeby się przekonać, jak nieobiektywne mogą być dane zaczerpnięte z portali społecznościowych, zastanówmy się nad relatywną popularnością szanowanego, przeznaczonego dla ambitnych odbiorców miesięcznika „Atlantic" i popularnością „National Enquirer", pisma

plotkarskiego i nierzadko żerującego na sensacji. Oba tytuły mają podobne nakłady[41], sprzedają się zaś w kilkuset tysiącach egzemplarzy („National Enquirer" to tygodnik, w sumie rozchodzi się więc nieco lepiej). W Google'u znajdziemy też porównywalną liczbę wyszukiwań obu tych czasopism.

Ale na Facebooku[42] blisko półtora miliona ludzi lajkuje „Atlantic" albo dyskutuje na swoich profilach o publikowanych w tym piśmie artykułach, a tylko pięćdziesiąt tysięcy polubia „Enquirera" bądź wymienia opinie o jego zawartości.

Popularność pism „Atlantic" i „National Enquirer" porównana na podstawie różnych źródeł

Nakład	mniej więcej jeden egzemplarz „Atlantica" przypada na jeden egzemplarz „National Enquirera"
Wyszukiwania w Google'u	jedno wyszukiwanie „Atlantica" przypada na jedno wyszukiwanie „National Enquirera"
Liczba polubień na Facebooku	dwadzieścia siedem polubień „Atlantica" przypada na jedno polubienie „National Enquirera"

Podstawą określenia popularności czasopisma jest jego nakład, co potwierdzają także dane z Google'a. Tymczasem dane z Facebooka przemawiają przytłaczająco na niekorzyść śmieciowego tabloidu i są najgorszym możliwym źródłem informacji, na podstawie których mielibyśmy rozstrzygać, co ludzie naprawdę lubią.

Z życiem jest tak samo, jak z naszymi preferencjami czytelniczymi. Na Facebooku prezentujemy się jako osoby kulturalne i nie odsłaniamy swojej prawdziwej twarzy. Dlatego używając danych z Facebooka w tej książce, a właściwie w tym rozdziale, zawsze pamiętam o powyższym zastrzeżeniu.

Żeby lepiej zrozumieć, czego nie widzą media społecznościowe, wróćmy jeszcze na chwilę do pornografii. W pierwszej kolejności powinniśmy się odnieść do powszechnego przekonania, że w Internecie dominują świństwa. To nieprawda. Większość treści w sieci nie jest pornograficzna. Na przykład wśród dziesięciu najczęściej odwiedzanych portali nie znajdziemy ani jednego pornograficznego[43]. A zatem chociaż popularność pornografii jest olbrzymia, nie należy jej przeceniać.

Mimo to jeżeli zbadać uważnie, jak lajkujemy pornografię i jak się nią dzielimy, to okaże się, że Facebook, Instagram i Twitter zaledwie uchylają okno na to, co jest faktycznie popularne w Internecie. Istnieją w sieci duże podzbiory stron cieszących się ogromnym zainteresowaniem, ale mało widocznych na społecznej scenie.

W chwili, gdy piszę te słowa, miano najpopularniejszego teledysku wszech czasów przypada *Gangnam Style* Koreańczyka PSY, kiczowatemu filmikowi do popowej piosenki, będącej satyrą na modnych Koreańczyków. Od debiutu w 2012 roku PSY tylko na YouTube miał dwa miliardy trzysta milionów odsłon, a jego sukces jest zupełnie oczywisty bez względu na to, jaki portal odwiedzasz, bo internauci dzielili się tym teledyskiem na różnych platformach społecznościowych dziesiątki milionów razy.

Z kolei najpopularniejszym w historii filmikiem porno jest zapewne *Great Body, Great Sex, Great Blowjob*, który oglądano ponad osiemdziesiąt milionów razy – czyli na trzydzieści odsłon *Gangnam Style* przypada przynajmniej jedna *Great Body, Great Sex, Great Blowjob*. Gdyby media społecznościowe przekazywały nam poprawny obraz popularności filmików i teledysków w Internecie, to jego użytkownicy powinni przesyłać sobie *Great Body, Great Sex, Great Blowjob* po milionkroć – lecz tym filmikiem dzielono się na portalach zaledwie kilkadziesiąt razy, a jego nadawcami i odbiorcami były bez wyjątku gwiazdy porno, nigdy zwyczajni ludzie, którzy najwyraźniej nie odczuwają potrzeby, aby chwalić się przed znajomymi, że interesują się tym utworkiem.

Facebook to cyfrowe serum typu „naplotę swoim znajomym, jak mi się fajnie żyje". W świecie Facebooka jego przeciętny dorosły użytkownik pozostaje w szczęśliwym związku małżeńskim, spędza

wakacje na Karaibach i przegląda „Atlantic". W realnym świecie ludzie w większości złoszczą się na wszystko, stoją w kolejkach do kasy w supermarketach, przeglądają „National Enquirera" i nie odbierają telefonów od swoich współmałżonków, z którymi nie sypiają już od wielu lat. W świecie Facebooka życie rodzinne wydaje się idealne – w rzeczywistości życie rodzinne to bagno. Czasami do tego stopnia, że niewielka grupa internautów żałuje nawet, że zdecydowali się mieć dzieci. W świecie Facebooka każdy młody dorosły baluje w sobotni wieczór na jakiejś ekstra imprze. W rzeczywistości większość z nich siedzi samotnie w domu i ogląda seriale na Netfliksie. W świecie Facebooka dziewczyna wysyła dwadzieścia sześć radosnych fotek z wycieczki ze swoim chłopakiem, a w świecie rzeczywistym natychmiast po ich wrzuceniu do sieci wpisuje w wyszukiwarkę „mój chłopak nie chce się ze mną kochać". I jest zupełnie możliwe, że w tym samym czasie jej chłopak ogląda w Internecie *Great Body, Great Sex, Great Blowjob*.

Cyfrowa prawda	**Cyfrowe kłamstwa**
Wyszukiwania	Posty w mediach społecznościowych
Odsłony	Lajki w mediach społecznościowych
Kliknięcia	Profile randkowe
Ściągnięcia	

PRAWDA O WASZYCH KLIENTACH

Wczesnym rankiem 5 września 2006 roku Facebook wprowadził ważne uaktualnienie na swoją stronę startową[44]. Pierwsze wersje Facebooka pozwalały klikać wyłącznie na profile znajomych, żebyśmy mogli się dowiedzieć, co porabiają. Portal, który jak sądzono, odniósł ogromny sukces, liczył wtedy dziewięć milionów czterysta tysięcy użytkowników.

Ale po kilku miesiącach ciężkiej pracy informatykom udało się stworzyć coś, co nazwali „aktualnościami" – funkcję, dzięki której

użytkownicy portalu dostawali najświeższe informacje o poczynaniach wszystkich swoich znajomych.

I ludzie korzystający z Facebooka natychmiast dali znać, że „aktualności" strasznie im się nie podobają. Ben Parr, student z Northwestern University, stworzył nawet internetową grupę „Studenci przeciwko «aktualnościom» na Facebooku". Jak twierdził, „«aktualności» są zbyt przerażające, stalkerowskie i muszą zniknąć". Kilka dni później grupa liczyła już siedemset tysięcy członków, podzielających opinię Parra. Pewien student drugiego roku z University of Michigan powiedział gazecie „Michigan Daily": „Nowy Facebook naprawdę mnie przeraża. Czuję się na nim jak stalker".

Jego wypowiedź przytacza David Kirkpatrick w swojej autoryzowanej historii portalu, *Efekt Facebooka*, w której określa wprowadzenie „aktualności" mianem „największego kryzysu, jaki kiedykolwiek dotknął Facebooka". Pisze jednak również, że kiedy przeprowadzał wywiad z Markiem Zuckerbergiem, współzałożycielem i szefem gwałtownie rozrastającej się firmy, ten wcale nie sprawiał wrażenia zmartwionego.

Dlaczego? Bo miał dostęp do cyfrowego serum prawdy: wiedział, jak często ludzie klikają i odwiedzają Facebooka. Kirkpatrick twierdzi, że Zuckerberg zdawał sobie sprawę z tego, że ludziom podobają się „aktualności" bez względu na to, co mówią o nich w podgrupach, i dysponował danymi, które tego dowodziły. Użytkownicy spędzali na Facebooku przeciętnie więcej czasu niż przed wprowadzeniem „aktualności". I stali się o wiele, wiele aktywniejsi. W sierpniu przejrzeli w serwisie dwanaście miliardów stron, z kolei w październiku, już po uruchomieniu „aktualności", dwadzieścia dwa miliardy.

Zuckerberg dysponował jednak jeszcze innymi dowodami, bo nawet zaraźliwa popularność grup występujących przeciwko „aktualnościom" wskazywała na siłę nowego serwisu. Grupa antyaktualnościowa rozrosła się tak szybko dokładnie dlatego, że wiele osób dowiedziało się, że przystąpili do niej ich znajomi – a ta wiadomość dotarła do nich właśnie za pośrednictwem „aktualności".

A zatem ludzie z jednej strony dołączali do potężnego chóru oburzonych i krzyczących, jacy to są nieszczęśliwi, że mogą znaleźć na Facebooku szczegółowe relacje z życia swoich znajomych, z drugiej zaś wracali na portal, żeby je przeczytać. „Aktualności" nie zostały zlikwidowane, a Facebooka odwiedza dzisiaj ponad miliard aktywnych użytkowników dziennie.

W książce *Zero to One* jej autor, Peter Thiel, jeden z pierwszych inwestorów Facebooka, powiada, że wielkie interesy robi się na sekretach – przyrody albo ludzkich[45]. Jeff Seder, wspomniany w rozdziale trzecim, odkrył tajemnicę lewej komory końskiego serca, dzięki której można przewidzieć, czy zwierzę będzie osiągać dobre wyniki na torze. A firma Google odkryła, jaką siłę mogą mieć informacje zawarte w linkach internetowych.

Thiel definiuje nasze sekrety jako coś, czego nie wiemy sami o sobie albo co ukrywamy, ponieważ nie chcemy, żeby wiedzieli o tym inni. Oznacza to, że wielkie interesy robi się na kłamstwach.

Można utrzymywać, że cały Facebook opiera się na niemiłej ludzkiej tajemnicy, którą Zuckerberg odkrył w czasie studiów na Harvardzie. W pierwszym semestrze drugiego roku nauki stworzył portal o nazwie Facemash. Wzorowany na portalu pod nazwą „Am I Hot or Not", Facemash przedstawiał zdjęcia dwóch studentów lub studentek Harvardu, po czym prosił kolegów i koleżanki o opinię, który/która z nich jest przystojniejszy/ładniejsza.

Portal młodego studenta wzbudził oburzenie. „Harvard Crimson" zarzucił Zuckerbergowi we wstępniaku, że odwołuje się do najgorszych stron natury ludzkiej, a Latynosi i Afroamerykanie oskarżyli go o seksizm i rasizm. Ale zanim harvardzcy administratorzy odcięli Zuckerbergowi dostęp do Internetu – zaledwie kilka godzin po otwarciu portalu – Facemasha odwiedziło czterysta pięćdziesiąt osób i oddało dwadzieścia dwa tysiące głosów na różne zdjęcia. W ten sposób Zuckerberg odkrył ważną tajemnicę: ludzie mogą twierdzić, że są wściekli, mogą coś potępiać i mówić, że to niesmaczne, ale będą klikać...

Dowiedział się też jeszcze jednej rzeczy: że chociaż internauci, w tym nawet studenci Harvardu, utrzymują, że są poważni, odpowiedzialni i szanują cudzą prywatność, to bardzo chętnie oceniają innych. Jego

wniosek wynikał z liczby odsłon i oddanych głosów. Później, ponieważ Facemash okazał się nazbyt kontrowersyjny, Zuckerberg, zorientowawszy się już, jak bardzo ciekawią nas trywialne wiadomości na temat osób, które rzekomo znamy, uczynił to odkrycie podwaliną jednego z największych przedsiębiorstw współczesności.

Na początku swojego istnienia nauczył się tego też Netflix: nie wierz w to, co ludzie mówią, tylko w to, co robią.

Firma początkowo pozwalała swoim klientom tworzyć coś w rodzaju listy filmów, które chcieliby obejrzeć w przyszłości, ponieważ na razie nie mają na to czasu. Dzięki temu Netflix mógł im później przypomnieć wybrane tytuły.

Ale pracownicy Netflixa zauważyli w danych coś dziwnego. Widzowie układali długie listy filmów, lecz kiedy platforma przypominała im o nich jakiś czas potem, rzadko w nie klikali.

Z czego to wynikało? Spytajcie użytkowników, co zamierzają obejrzeć w najbliższych dniach, a wypełnią listę ambitnymi, wyrafinowanymi filmami w rodzaju czarno-białych dokumentów o drugiej wojnie światowej albo poważnych dzieł zagranicznych. Po paru dniach będą jednak oglądali to samo, co zwykle: kiczowate komedie lub romanse. Ludzie notorycznie okłamują samych siebie.

W obliczu takiego rozdźwięku Netflix przestał prosić widzów, żeby pisali, co chcą zobaczyć w przyszłości, i przygotował ofertę na podstawie milionów kliknięć i odsłon klientów podobnych platform kinowych. Firma zaczęła więc witać użytkowników sugerowaną listą filmów, utworzoną nie na podstawie tego, co rzekomo podobało się widzom, ale na podstawie danych wskazujących, co najprawdopodobniej rzeczywiście obejrzą. Rezultat: klienci odwiedzali Netflixa częściej i oglądali więcej filmów.

„Algorytmy znają cię lepiej, niż ty znasz sam siebie", powiada Xavier Amatriain, były netfliksowy analityk danych[46].

Dlaczego nie warto zwracać uwagi na to, co ludzie mówią

Co ludzie mówią	**Rzeczywistość**	**Ipso facto**
Nie chcą stalkować znajomych	Mało czego na świecie pragną tak bardzo, jak oceniać znajomych i wiedzieć, co robią	Mark Zuckerberg, współzałożyciel Facebooka, jest wart 55,2 miliarda dolarów
Nie chcą kupować towarów będących wytworem wyzysku taniej siły roboczej	Kupują ładne produkty „w rozsądnej cenie”	Phil Knight, współzałożyciel Nike, jest wart 25,4 miliarda dolarów
Rano chcą słuchać wiadomości	Rano chcą słuchać o karłach i karlicach uprawiających seks z gwiazdami porno	Howard Stern jest wart 500 milionów dolarów
Nie chcą czytać o bondage, dominacji i sadomasochizmie	Chcą czytać o seksie, BDSM między młodą absolwentką college’u i magnatem biznesowym	*Pięćdziesiąt twarzy Greya* sprzedało się w 125 milionach egzemplarzy
Chcą, żeby politycy jasno określali swoje poglądy polityczne	Chcą, żeby politycy oszczędzali im szczegółów, ale by sprawiali wrażenie twardych i pewnych siebie	Donald Trump

CZY POTRAFIMY SIĘ POGODZIĆ Z PRAWDĄ?

Być może niektóre fragmenty tego rozdziału wprawiają was w przygnębienie. Cyfrowe serum prawdy ujawniło, że bardzo lubimy oceniać ludzi po wyglądzie, że swoją orientację seksualną wciąż ukrywają przed światem miliony gejów, że znaczny odsetek kobiet fantazjuje o gwałcie, że w USA panuje powszechna niechęć do Afroamerykanów, że chociaż nie pokazują tego statystyki, to wiele osób stosuje przemoc wobec dzieci i własnoręcznie przeprowadza zabiegi przerywania ciąży, i że obserwujemy gwałtowny wybuch islamofobii, która przybrała tylko na sile, kiedy prezydent zaapelował o więcej tolerancji. Nie są to budujące

wiadomości. Często się zdarza, że gdy przedstawiam swoje badania na wykładach, po ich zakończeniu podchodzą do mnie słuchacze i mówią: „Wszystko to brzmi szalenie interesująco, Seth. Ale jest okropnie przygnębiające".

Nie mogę udawać, że w części tych danych nie kryje się coś ponurego. Jeżeli ludzie nieustannie mówią nam to, co – ich zdaniem – chcemy usłyszeć, to najczęściej będą mówić rzeczy bardziej krzepiące niż prawda. Cyfrowe serum prawdy na ogół pokazuje, że świat jest gorszy, niż nam się wydaje.

Czy naprawdę warto być tego świadomym? Przecież gdy poznajemy treści wpisów w Google'u, dane z portali pornograficznych, albo dowiadujemy się, kto w co klika, nie zawsze myślimy w duchu: „To wspaniale. Wreszcie możemy zrozumieć, kim naprawdę jesteśmy", tylko raczej: „To straszne. Wreszcie możemy zrozumieć, kim naprawdę jesteśmy".

W ostatecznym rozrachunku ta prawda jednak pomaga – nie tylko Markowi Zuckerbergowi czy innym przedsiębiorcom, którzy starają się zwiększyć „klikalność" swoich stron albo przyciągnąć klientów. Wiedza płynąca z analizy danych może nam pomóc w życiu co najmniej na trzy sposoby.

Na przykład możemy czerpać pociechę z tego, że nie jesteśmy sami ze swoją niepewnością i zawstydzającym zachowaniem. Przyjemnie jest wiedzieć, że inni ludzie też mają zastrzeżenia do swojego wyglądu. Wiele osób, zwłaszcza tych, które nie kochają się zbyt często, dowie się z przyjemnością, że nie wszyscy na świecie parzą się bez przerwy jak króliki. A wreszcie może jakiemuś licealiście z Missisipi, podkochującemu się w rozgrywającym drużyny futbolowej, pomoże świadomość, że chociaż otaczają go tylko nieliczni geje otwarcie przyznający się do swojej orientacji, to wielu innych mężczyzn ma takie problemy jak on.

Istnieje jeszcze jedna sfera, o której dotąd nie wspominałem, a gdzie wpisy w Google'u mogą nam pomóc zrozumieć, że nie jesteśmy sami ze swoimi problemami. Zapewne wszystkim nam mówiono w szkole, że jeśli mamy pytanie, to trzeba podnieść rękę, bo skoro czegoś nie rozumiemy, to najprawdopodobniej nie rozumieją tego także inni

uczniowie. Jeżeli jednak byliście podobni do mnie, to nie słuchaliście nauczyciela, tylko siedzieliście cicho, bojąc się otworzyć usta. Myśleliście, że wasze pytanie okaże się za głupie – wszyscy wydawali się wam mądrzejsi. Anonimowy zbiór danych z wyszukiwarki demonstruje jednak raz na zawsze, że nauczyciele mieli bezsprzeczną rację; w cudzych głowach też często lęgną się podstawowe i wcale niekoniecznie mądre pytania.

Zobaczmy więc, o co najczęściej pytali Amerykanie w czasie orędzia o stanie państwa, wygłoszonego przez prezydenta Obamę w 2014 roku[47].

Nie tylko wy się nad tym zastanawiacie: oto cztery pytania najczęściej zadawane w Internecie w czasie orędzia o stanie państwa

Ile lat ma Barack Obama?
Kto siedzi koło Bidena?
Dlaczego Boehner jest w zielonym krawacie?
Dlaczego Boehner jest pomarańczowy?

Czytając powyższe pytania, można pomyśleć, że nie świadczą one najlepiej o naszej demokracji. To, że bardziej niż treścią przemówienia interesujemy się kolorem czyjegoś krawata albo skóry, nie stawia nas w dobrym świetle. Nie wiedzieć, kim jest John Boehner, wówczas przewodniczący Izby Reprezentantów, też nie dowodzi naszego zaangażowania w politykę wewnętrzną.

Dlatego traktuję powyższe wyszukiwania jako dowód na mądrość naszych nauczycieli. Zwykle nie zadajemy takich pytań jak te, bo zdają się nam za głupie. Wielu osobom przychodzą jednak do głowy – więc wpisują je w Google'a.

Uważam nawet, że big data może pomóc nam uaktualnić takie powiedzenie: „Nigdy nie stawiaj znaku równości pomiędzy tym, co myślisz, i tym, co ludzie mówią". Jego dwudziestopierwszowieczna wersja spod znaku big data mogłaby brzmieć tak: „Nigdy nie stawiaj

znaku równości pomiędzy swoimi wyszukiwaniami w sieci i postami w mediach społecznościowych".

Spójrzcie na przykład, jak kobiety przedstawiają swoich mężów na portalach, a jak w anonimowych wyszukiwaniach.

Pięć najczęściej spotykanych opisów mężów

Posty w mediach społecznościowych	**Wyszukiwania**
najlepszy	gej
mój najlepszy przyjaciel	palant
wspaniały	wspaniały
najcudowniejszy	irytujący
przeuroczy	skąpiec

Widzimy posty w mediach społecznościowych, ale nie widzimy wyszukiwań, wydaje się nam więc, że mnóstwo pań stale określa swoich małżonków mianem „najlepszych", „największych" i „przeuroczych"[13*], i mamy skłonności do tego, aby minimalizować liczbę kobiet, które uważają swoich mężów za „palantów", „irytujących" czy „skąpców". Poddając analizie anonimowe dane, wszyscy możemy zrozumieć, że nie tylko nam życie i małżeństwo wydają się trudne. I może się nareszcie nauczymy, że nie należy porównywać naszych wyszukiwań do postów w *social media*.

Druga korzyść z cyfrowego serum prawdy jest taka, że zwraca naszą uwagę na cierpiących. Organizacja Human Rights Campaign poprosiła, bym podjął z nią współpracę i pomógł przekonać homoseksualnych mężczyzn w niektórych stanach, że mogą ujawnić swoją orientację. HRC chce wykorzystać anonimowe dane z Google'a do tego, by zdecydować, gdzie takie osoby w pierwszej kolejności powinny się uaktywnić. Kontaktowały się ze mną również organizacje opiekuńcze i pomocy dzieciom, żeby się dowiedzieć, w jakich częściach kraju może

dochodzić do aktów przemocy częściej, niż wynikałoby to z liczby zgłoszeń.

Podsunięto mi jeszcze jeden zaskakujący temat: zapachów waginalnych. Gdy pierwszy raz pisałem o tej sprawie, i to nie byle gdzie, bo w samym „New York Timesie", zrobiłem to dla żartu. I ja, i moi koledzy podśmiewaliśmy się trochę z tego felietonu.

Kiedy jednak zajrzałem potem na fora internetowe, które pojawiają się czasem w wynikach wyszukiwań na podobne tematy, okazało się, że pisze na nich mnóstwo młodych dziewcząt, przekonanych, że brzydki zapach z pochwy zatruwa im życie. Dlatego skontaktowali się ze mną eksperci od wychowania seksualnego z pytaniem, w jaki sposób mogliby wykorzystać chociaż część danych internetowych, by zmniejszyć ów paranoiczny lęk wśród młodych dziewczyn.

Nie czuję się wprawdzie szczególnie kompetentny w żadnej z powyższych kwestii, są one jednak poważne i uważam, że analiza danych może pomóc je rozwiązać.

Ostatnią – i moim zdaniem najpotężniejszą – wartością cyfrowego serum prawdy jest jego zdolność do prowadzenia nas od problemów ku ich rozwiązaniom. Gdy lepiej zrozumiemy naturę ludzką, może uda się nam ograniczyć podłość i zminimalizować zło na świecie.

Wróćmy jednak do przemówienia Obamy na temat islamofobii. Przypomnijmy sobie, że za każdym razem, gdy prezydent mówił, że powinniśmy bardziej szanować muzułmanów, ludzie, do których usiłował trafić, wpadali w coraz większą złość.

Ale wyszukiwania w Google'u ujawniają, że Obama przynajmniej w jednym zdaniu powiedział coś, co wywołało taką reakcję, jakiej mógłby sobie życzyć: „Amerykańscy muzułmanie to nasi przyjaciele, sąsiedzi, koledzy z pracy, doskonali sportowcy, a także, tak, żołnierze i żołnierki w mundurach, gotowi umierać w obronie naszej ojczyzny".

Po tej wypowiedzi pierwszy raz od ponad roku najczęściej przywoływanym słowem po „muzułmańscy" nie było „terroryści", „ekstremiści" czy „uchodźcy", ale „sportowcy", a następnie „żołnierze". A „sportowcy" utrzymywali się na pierwszym miejscu jeszcze przez cały następny dzień.

Dane internetowe sugerują, że kiedy prawimy kazania komuś, kto się złości, jego gniew może tylko rosnąć. Ale jeżeli będziemy subtelnie pobudzać ciekawość, podawać nowe informacje i podsuwać ludziom nowe wizerunki grup społecznych, które budzą ich niechęć, myśli słuchaczy mogą zacząć biec w innych, bardziej pozytywnych kierunkach.

Dwa miesiące później Obama, tym razem w meczecie, wygłosił kolejne transmitowane przez telewizję przemówienie o islamofobii[48]. Być może ktoś z gabinetu prezydenta przeczytał felietony moje i Soltasa w „Timesie", gdzie napisaliśmy, co naszym zdaniem wcześniej odniosło pożądany skutek, a co nie, bo treść drugiego wystąpienia Obamy różniła się wyraźnie od pierwszego.

Tym razem prezydent poświęcił niewiele czasu omówieniu wartości tolerancji, skupił się natomiast na pobudzeniu ciekawości odbiorców i zmianie naszego sposobu postrzegania amerykańskich muzułmanów. Podkreślił między innymi, że wielu niewolników z Afryki wyznawało islam, że Thomas Jefferson i John Adams mieli w swoich bibliotekach egzemplarze Koranu, że pierwszy meczet na amerykańskiej ziemi wzniesiono w Dakocie Północnej, a kilka wieżowców w Chicago projektował muzułmański architekt. Powtórzył również to, co mówił o islamskich sportowcach i żołnierzach, dodał jednak przy tym, że mamy muzułmanów w policji i straży pożarnej, a także wśród nauczycieli i lekarzy.

Moja analiza wyszukiwań w Google'u wskazuje, że to przemówienie było znacznie skuteczniejsze. Tym razem w ciągu kilku godzin po wystąpieniu Obamy liczba nienawistnych i pełnych wściekłości wpisów przeciwko muzułmanom znacznie spadła[49].

Istnieją jeszcze inne potencjalne sposoby wykorzystania danych z wyszukiwarki, by sprawdzić, co wywołuje lub zmniejsza naszą nienawiść. Spójrzmy na przykład, jak zmieniają się rasistowskie wpisy internetowe, kiedy miejscowa drużyna futbolowa pozyska czarnego rozgrywającego, albo jak ewoluują wpisy seksistowskie, kiedy na jakiś ważny urząd zostaje wybrana kobieta. Możemy też zobaczyć, w jaki sposób reaguje rasizm na tak zwaną policję środowiskową, a seksizm na nowe prawa o molestowaniu seksualnym.

Przydatna może okazać się również wiedza o naszych podświadomych uprzedzeniach. Moglibyśmy na przykład trochę się wysilić i bardziej zachwycać intelektem dziewcząt, niż martwić ich urodą. Zbiory danych z Google'a i inne internetowe źródła prawdy dają nam bezprecedensowy wgląd w najmroczniejsze zakamarki ludzkiej duszy. Przyznaję, że z tą nową wiedzą trudno jest się czasem pogodzić, ale wraz z nią dostajemy też moc, którą możemy wykorzystywać do walki z siłami ciemności. A zgromadzenie dużego zbioru danych o najważniejszych dla świata problemach to pierwszy krok na drodze do ich wyeliminowania.

ROZDZIAŁ PIĄTY
ZBLIŻENIE

Mój brat, Noah, jest ode mnie cztery lata młodszy. Ludzie, którzy widzą nas razem po raz pierwszy, uważają, że jesteśmy niesamowicie podobni. Obaj zbyt głośno mówimy, łysiejemy w identyczny sposób i mamy poważne problemy z utrzymaniem porządku w domu.

Są jednak i różnice – ja liczę każdy grosz, a Noah kupuje wszystko, co najdroższe. Ja uwielbiam Leonarda Cohena i Boba Dylana, a Noah zespół Cake i Becka.

Być może najlepiej widoczną różnicą między nami jest nasz stosunek do bejsbolu. Ja mam obsesję na jego punkcie i zwłaszcza moja miłość do nowojorskich Metsów zawsze stanowiła część mojej tożsamości. Z kolei Noah uważa, że bejsbol jest niewiarygodnie nudny, a częścią jego tożsamości od wczesnego dzieciństwa była niechęć do tej dyscypliny sportu[14*].

Seth Stephens-Davidowitz
bejsbolofil

Noah Stephens-Davidowitz
bejsbolofob

Jak to możliwe, że dwóch facetów o podobnych genach, wychowanych przez tych samych rodziców i w tym samym mieście, może żywić tak krańcowo sprzeczne uczucia do bejsbolu? Co sprawia, że wyrastamy na takich, a nie innych ludzi? A mówiąc jeszcze prościej, co jest nie w porządku z moim bratem?

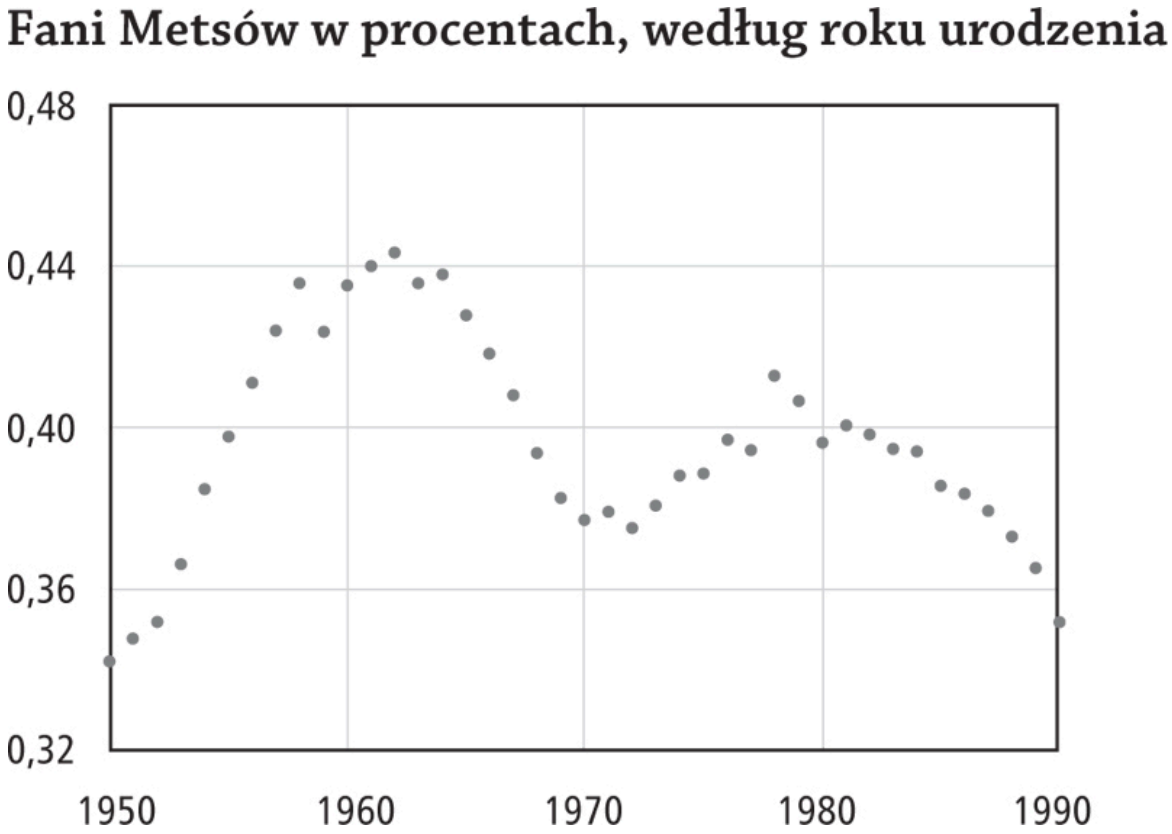
W psychologii rozwoju powstała niedawno nowa gałąź, której przedstawiciele gromadzą gigantyczne bazy danych na temat osób dorosłych i usiłują skorelować je z najważniejszymi zdarzeniami z dzieciństwa. Ta metoda pozyskiwania wiedzy może pomóc nam odpowiedzieć na postawione wyżej i podobne pytania. A coraz częstsze wykorzystywanie big data do rozwiązywania problemów psychologicznych możemy nazywać big psych.

By się przekonać, w czym rzecz, przyjrzyjmy się przeprowadzonym przeze mnie badaniom, sprawdzającym, w jaki sposób nasze doświadczenia z dzieciństwa wpływają na to, której drużynie bejsbolowej kibicujemy – i czy w ogóle mamy ulubioną drużynę[1]. Wykorzystałem w tym celu „lajki" dla drużyn bejsbolowych na Facebooku (w poprzednim rozdziale zauważyłem, że dane z portali społecznościowych mogą być bardzo mylące w pewnych wrażliwych sprawach, tutaj zakładam jednak, że nikt, nawet kibic Phillies, nie

wstydzi się przyznać na Facebooku, że zaczyna się interesować jakąś drużyną).

Zacząłem od tego, że zapisałem sobie na twardym dysku, ilu mężczyzn (w każdym wieku) „lubi" obie nowojorskie drużyny bejsbolowe. Oto odsetek fanów Metsów, pogrupowanych wedle roku urodzenia:

Fani Metsów w procentach, według roku urodzenia

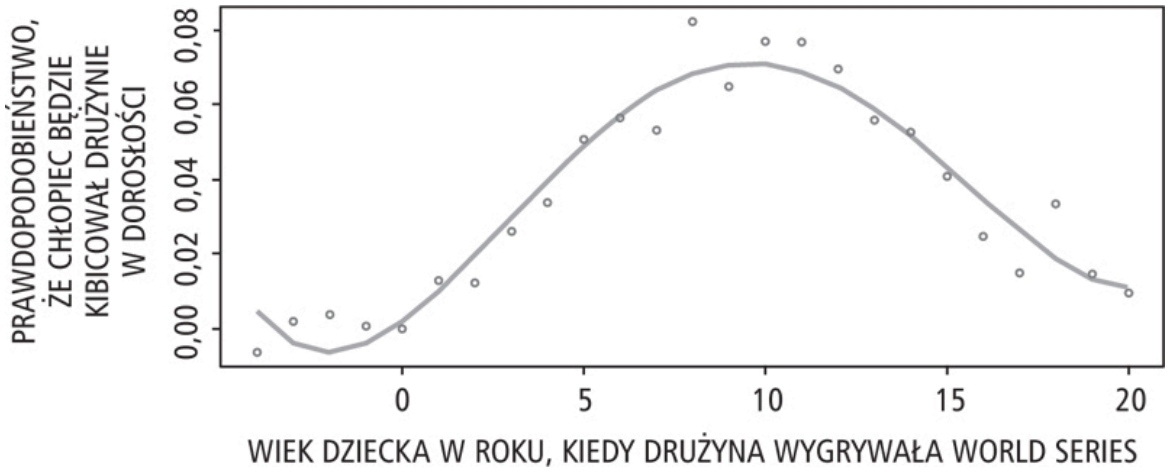


Im wyższy punkt na wykresie, tym więcej fanów. Popularność zespołu rośnie i spada, potem jeszcze raz, a największym zainteresowaniem Metsi cieszą się wśród mężczyzn urodzonych między 1962 i 1978 rokiem. Domyślam się, że wielbiciele bejsbolu wiedzą już, dlaczego: bo Metsi tylko dwa razy zwyciężyli w World Series – w roku 1969 i 1986, co znaczy, że ci ich wspomniani wyżej kibice mieli wtedy po siedem lub osiem lat. Można zatem powiedzieć, że mężczyźni

najczęściej zostawali fanami Metsów jako siedmio- albo ośmiolatkowie, kiedy ich drużyna wygrywała World Series.

Tę analizę da się kontynuować. W następnej kolejności zapisałem na dysk informacje z Facebooka, by sprawdzić, ilu fanów (w każdym wieku) „lubi" dowolne drużyny grające w pierwszej lidze bejsbolowej (Major League).

I przekonałem się między innymi, że istnieje też niezwykle duża grupa kibiców drużyny Baltimore Orioles, urodzonych w 1962 roku, oraz zwolenników Pittsburgh Pirates, urodzonych w roku 1963 – ci fani również mieli po osiem lat, kiedy ich ulubione drużyny zdobywały mistrzostwo. Obliczyłem następnie, w jakim wieku były te osoby, gdy zespoły, którym kibicowały, uzyskiwały największą popularność, i przedstawiłem wyniki na takim wykresie:



I w tym wypadku widać, że dla mężczyzny moment decydujący o tym, czy będzie kibicował danej drużynie także w wieku dorosłym, przypada mniej więcej na ósmy rok jego życia – a najogólniej rzecz biorąc, zespół zdobywa sobie najwięcej sympatyków wśród chłopców w wieku od pięciu do piętnastu lat. Zwycięstwo w World Series, gdy kibic ukończył już lat dziewiętnaście albo dwadzieścia, ma około ośmiu razy mniejsze znaczenie dla tego, czy pozostanie wiernym fanem drużyny, niż jeśli zdobyła mistrzostwo, kiedy był ośmiolatkiem. Więc gdy masz osiem lat, wiadomo już, czy będziesz uwielbiał swój zespół przez całe życie, czy nie.

Możecie zapytać, czy sprawa ma się podobnie w wypadku fanek. Prawidłowości wśród kobiet są znacznie mniej wyraźne, wygląda jednak na to, że większość pań decyduje się na wspieranie takiej czy innej drużyny, kiedy kończy dwudziesty drugi rok życia.

Ze wszystkich swoich analiz tę właśnie lubię najbardziej, gdyż wiąże się ona z dwoma moimi ulubionymi tematami badawczymi: bejsbolem i źródłem wewnętrznego niezadowolenia w dorosłym życiu. Wiernym kibicem zostałem w roku 1986 i od tamtej pory wciąż cierpię, wspierając Metsów. Noah miał więcej oleju w głowie, urodził się bowiem cztery lata później i bóle porażek zostały mu oszczędzone.

Oczywiście bejsbol to nie najważniejsza rzecz na świecie, a w każdym razie tak powtarzali promotorzy mojej pracy doktorskiej. Ale stosowana przeze mnie metoda bejsbolowa może nam pomóc uzyskać też odpowiedzi na podobne pytania, w tym – jak i kiedy powstają nasze preferencje polityczne, upodobania seksualne, gusta muzyczne oraz przyzwyczajenia finansowe (interesuje mnie zwłaszcza, skąd wzięły się porąbane poglądy mojego brata w tych dwóch ostatnich kwestiach). Ośmielam się twierdzić, że wiele naszych zachowań i zainteresowań, nawet takich, które uważamy za fundamentalne dla naszej osobowości, można wyjaśnić pewnymi losowymi zdarzeniami z dzieciństwa oraz tym, co działo się na świecie w kluczowych dla naszego rozwoju wczesnych latach życia.

Tą metodą przeprowadzono już nawet pewne badania dotyczące źródeł preferencji politycznych. Yair Ghitza, szef zespołu naukowców w Catalist, firmie zajmującej się analizą danych, i Andrew Gelman, politolog i statystyk z Columbia University, postanowili sprawdzić rozpowszechnione mniemanie, że większość z nas zaczyna jako liberałowie, a w konserwatyzm popada stopniowo z wiekiem. Pogląd ten wyraża się w przypisywanym często Winstonowi Churchillowi powiedzeniu, głoszącym, że każdy człowiek przed trzydziestką, który nie jest liberałem, nie ma serca, a każdy człowiek po trzydziestce, który nie jest konserwatystą, nie ma rozumu.

Ghitza i Gelman przeanalizowali dane z ostatnich sześćdziesięciu lat i przyjrzeli się preferencjom wyborczym przeszło trzystu tysięcy osób. Wbrew powiedzeniu Churchilla okazało się, że nastoletnia młodzież

zdradza czasami skłonności liberalne, a czasem konserwatywne, podobnie jak ludzie w średnim wieku i starsi.

Badacze odkryli też, że nasze sympatie światopoglądowe rodzą się w podobny sposób jak miłość do określonych klubów sportowych. Jest pewien okres w życiu, który ukształtowuje nas już na zawsze. Na przykład wielu Amerykanów w wieku od czternastu do dwudziestu czterech lat przyjmuje poglądy polityczne wedle stopnia popularności urzędującego w tym czasie prezydenta. Popularny republikanin albo niepopularny demokrata może sprawić, że wielu młodych dorosłych będzie w przyszłości wspierać republikanów, a z kolei niepopularny republikanin albo popularny demokrata przysporzy zwolenników – łatwo ulegających wpływom zewnętrznym – demokratom.

Inklinacje powstałe w kluczowych dla tej kwestii latach życia zostają z nami do śmierci. Zobaczmy, jak to wygląda na przykładzie Amerykanów urodzonych w roku 1941 oraz dziesięć lat później.

Osoby z tej pierwszej grupy weszły w dorosłość za prezydentury Dwighta D. Eisenhowera, popularnego republikanina. Na początku lat sześćdziesiątych pokolenie roku 1941 mocno popierało Partię Republikańską, choć przecież nie osiągnęło jeszcze nawet trzydziestki, i z wiekiem konsekwentnie podtrzymywało swoją sympatię dla republikanów.

Z kolei Amerykanie urodzeni dziesięć lat później – pokolenie „baby boom" – dorastali za czasów prezydentury Johna F. Kennedy'ego, niebywale popularnego demokraty, Lyndona B. Johnsona, demokraty popularnego na początku swojego urzędowania, i Richarda M. Nixona, republikanina, który okryty niesławą, został ostatecznie zmuszony do opuszczenia Białego Domu. Przedstawiciele tej generacji konsekwentnie reprezentowali poglądy liberalne.

Mając do dyspozycji wszystkie te dane, uczeni wskazali najważniejszy rok życia dla formowania się naszych poglądów politycznych: osiemnasty[2].

Okazało się bowiem, że siła imprintingu z tego czasu jest znaczna. Z obliczeń Ghitzy i Gelmana wynika, że wśród Amerykanów urodzonych w 1941 roku „doświadczenie Eisenhowera" przyniosło

wynoszący dziesięć punktów procentowych wzrost poparcia dla Partii Republikańskiej, a „doświadczenie Kennedy'ego, Johnsona i Nixona" dało demokratom przewagę siedmiu punktów procentowych wśród Amerykanów urodzonych w roku 1952.

Powiedziałem już jasno, że podchodzę sceptycznie do danych sondażowych, w tym wypadku jestem jednak pod wrażeniem liczby pozyskanych odpowiedzi. Takiego badania nie dałoby się zresztą przeprowadzić za pomocą zaledwie jednego małego sondażu. Uczeni potrzebowali setek tysięcy odpowiedzi, pochodzących z wielu ankiet, by adekwatnie zbadać, jak z wiekiem zmieniają się nasze upodobania.

Wielkość zbioru danych była też zasadnicza dla mojej analizy bejsbolowej. Musiałem przypatrzyć się z bliska nie tylko fanom poszczególnych drużyn, ale też osobom ze wszystkich grup wiekowych. Potrzebne są do tego miliony komentarzy i wpisów, a Facebook i inne cyfrowe źródła informacji dostarczają nam ich w takiej ogromnej ilości regularnie.

Wielkość big data zaczyna liczyć się właściwie dopiero w tym punkcie. Żeby dobrze się przyjrzeć niewielkiemu fragmentowi zdjęcia, potrzeba wielu pikseli, tak jak potrzeba wielu danych w zbiorze, by dobrze przyjrzeć się jego niewielkiemu podzbiorowi i na przykład określić, jaką popularnością cieszą się Metsi wśród mężczyzn urodzonych po 1978 roku. W sondażu na małą skalę, obejmującym kilka tysięcy osób, nie znajdzie się odpowiednio liczna grupa takich respondentów.

Wracamy tu do trzeciej mocy big data – ponieważ dysponujemy ogromnym zbiorem danych, możemy przypatrzyć się z bliska jego niewielkim podzbiorom i dzięki temu lepiej poznać siebie. I oczywiście możemy wziąć pod lupę inne aspekty niż wiek. Jeśli będziemy mieli odpowiednio duży zbiór, uda nam się zobaczyć, jak zachowują się ludzie w poszczególnych miastach i miasteczkach, godzina po godzinie, a nawet minuta po minucie.

Dlatego w tym rozdziale zrobimy sobie zbliżenie na rozmaite ludzkie postawy i zachowania.

CO NAPRAWDĘ DZIEJE SIĘ W AMERYKAŃSKICH HRABSTWACH, MIASTACH I MIASTECZKACH?

Z perspektywy czasu wydaje się to może zaskakujące, ale kiedy Raj Chetty, wówczas profesor Harvardu, i jego niewielki zespół badawczy po raz pierwszy dostali do ręki całkiem pokaźny zbiór danych – ewidencję podatkową wszystkich Amerykanów, począwszy od 1996 roku – wcale nie byli pewni, czy ich badania przyniosą jakiekolwiek rezultaty. Urząd skarbowy (IRS) przekazał im te dane, sądząc, że uczeni pomogą mu określić efektywność polityki podatkowej prowadzonej przez państwo.

Pierwsze próby, jakie podjął Chetty ze swoim zespołem, rzeczywiście zabrnęły w ślepą uliczkę, bo ich analiza na poziomie kraju i poszczególnych stanów prowadziła w większości do takich samych wniosków, jakie można było wyciągnąć na podstawie sondaży, choć może wyniki uzyskane przez Chetty'ego były – dzięki setkom milionów tak zwanych punktów danych ze skarbówki – trochę dokładniejsze. Ale uzyskanie identycznej odpowiedzi jak wszyscy, choćby nawet bardziej precyzyjnej, nie jest wielkim osiągnięciem w dziedzinie nauk społecznych, a poważne czasopisma naukowe nie palą się do publikacji tego rodzaju odkryć.

Poza tym organizacja i analiza całego zbioru danych z IRS była czasochłonna. Chetty i jego zespół, zasypani lawiną informacji, potrzebowali więcej czasu na to, żeby uzyskać takie same odpowiedzi jak wszyscy inni badacze.

Wydawało się więc, że naukowcy nastawieni sceptycznie do big data mieli rację. Nie potrzeba danych o kilkuset milionach Amerykanów, żeby zrozumieć politykę podatkową – do tego celu w pełni wystarczy sondaż obejmujący dziesięć tysięcy osób. Jest zatem zrozumiałe, że Chetty i członkowie jego zespołu czuli się zniechęceni.

I dopiero wtedy zrozumieli swój błąd. „Nie chodzi o to, żeby robić to samo, co z wynikami sondaży, bo nie jest tak, że big data niczym się od nich nie różni i tylko dostarcza po prostu więcej informacji", tłumaczy Chetty, który wraz z zespołem poszukiwał odpowiedzi na pytania dotyczące wyłącznie drobnych fragmentów wielkiego zbioru danych[3].

„Big data tak naprawdę pozwala nam korzystać z zupełnie innych metod badawczych niż sondaż – dodaje. – Możemy na przykład przyjrzeć się bliżej detalom geograficznym".

To znaczy, że dysponując informacjami o wielu milionach osób, Chetty i jego ludzie mogli zaobserwować pewne prawidłowości w miastach, miasteczkach i osiedlach, zarówno dużych, jak i małych.

Jako seminarzysta na Harvardzie siedziałem na sali wykładowej, gdy Chetty przedstawiał wstępne wyniki swoich badań, prowadzonych na podstawie ewidencji podatkowej wszystkich obywateli Stanów Zjednoczonych. Socjologowie mówią w swoim żargonie o „obserwacjach" – czyli o tym, ile uzyskują punktów danych. Socjolog pracujący na podstawie wyników sondażu, w którym odpowiadało osiemset osób, powie, że ma „osiemset obserwacji", a jeżeli prowadzi eksperyment obejmujący siedemdziesiąt osób, to powie, że ma tych obserwacji siedemdziesiąt.

„Mamy miliard dwieście milionów obserwacji", stwierdził bez zmrużenia oka Chetty, a słuchacze zachichotali nerwowo.

Wtedy, na tej sali, a później w cyklu artykułów naukowych, Chetty i współautorzy jego badań zaczęli nam pokazywać nowy obraz Ameryki.

Rozważmy takie pytanie: czy Stany Zjednoczone są krajem nieograniczonych możliwości? Czy jeżeli nie masz bogatych rodziców, istnieje jakaś szansa, że wzbogacisz się o własnych siłach?

Tradycyjny sposób poszukiwania odpowiedzi polegałby na tym, że przyjrzelibyśmy się reprezentatywnej grupie Amerykanów i porównalibyśmy ją z podobną grupą osób w innych krajach.

Poniżej podaję dane dla kilku państw, pokazujące możliwości odniesienia indywidualnego sukcesu finansowego. Pytanie brzmi: jakie są szanse, że osoba urodzona w rodzinie należącej do dwudziestu procent mieszkańców o najniższych poborach osiągnie w przyszłości dochody plasujące ją wśród dwudziestu procent najzamożniejszych obywateli?

Szanse, że dziecko ubogich rodziców się wzbogaci (cztery wybrane kraje)

Stany Zjednoczone	7,5
Wielka Brytania	9,0
Dania	11,7
Kanada	13,5

Jak widzimy, USA wcale nie wypadają za dobrze w tym zestawieniu.

Taka prosta analiza nie pokazuje jednak, jak jest naprawdę. Bo zespół Chetty'ego przyjrzał się też różnicom geograficznym – i wtedy okazało się, że mamy bardzo różne szanse w zależności od tego, w jakim rejonie kraju przyszliśmy na świat.

Szanse, że dziecko ubogich rodziców się wzbogaci (wybrane miasta Stanów Zjednoczonych)

San Jose, Kalifornia	12,9
Waszyngton, Dystrykt Kolumbii	10,5
średnia dla USA	7,5
Chicago, Illinois	6,5
Charlotte, Karolina Północna	4,4

W niektórych rejonach Stanów Zjednoczonych dziecko biednych rodziców ma zatem równie wysokie szanse na sukces jak w najbardziej rozwiniętych krajach świata – gdzie indziej jednak te szanse są znacznie niższe.

Tych prawidłowości nie zobaczylibyśmy nigdy w wynikach małego sondażu, w którym wzięłoby udział co najwyżej kilku ankietowanych z Charlotte i San Jose, nie moglibyśmy więc przyjrzeć się niektórym amerykańskim miastom jak na filmowym zbliżeniu.

Ale zespół Chetty'ego był w stanie przeprowadzić jeszcze dokładniejsze badania. Ponieważ dysponował ogromnym zbiorem danych, dotyczących każdego Amerykanina, mógł przyjrzeć się nawet nielicznym grupom osób, które przeprowadzały się z miasta do miasta, i sprawdzić, w jaki sposób zmieniało to ich szanse. Chodziło na przykład o ludzi, którzy przeprowadzali się z Nowego Jorku do Los Angeles, z Milwaukee do Atlanty czy z San Jose do Charlotte. Dzięki temu mogli też badać przyczyny, nie tylko korelacje (omówię to rozróżnienie w następnym rozdziale). Okazało się, że owszem, przeprowadzka do odpowiedniego miasta w latach, kiedy kształtuje się nasza osobowość, czyni znaczną różnicę.

Czy zatem Ameryka jest krajem nieograniczonych możliwości?

Odpowiedź brzmi: ani tak, ani nie. A raczej: w pewnych miejscach tak, a w pewnych miejscach nie.

Jak piszą autorzy, najlepiej jest przedstawić Stany Zjednoczone jako zespół społeczności, z których część stwarza szanse sukcesu jednostkom i już od kilku pokoleń umożliwia im znaczny wzrost dochodów, część zaś pozwala uniknąć ubóstwa tylko nielicznym.

Co decyduje o tym, że w pewnych regionach kraju dochody mieszkańców znacznie rosną? Dlaczego w niektórych miejscach szanse zostają wyrównane i biedny dzieciak może tam prowadzić naprawdę niezłe życie? Otóż lepsze pod tym względem są te miasta czy hrabstwa, które wydają więcej na edukację, oraz te zamieszkane przez większą liczbę osób religijnych i gdzie jest niższa przestępczość. Gorzej natomiast wypadają regiony, gdzie większość populacji stanowią Afroamerykanie – i co ciekawe, szanse sukcesu są tam mniejsze nie tylko dla czarnoskórych, ale też dla białych dzieci. Równie nisko plasują się te miejsca, gdzie mieszka dużo samotnych matek, a szanse na sukces okazują się tam skromniejsze nie tylko dla ich dzieci, ale i dla dzieci z pełnych rodzin. Niektóre wyniki sugerują, że ważne jest to, w jakim otoczeniu wychowujemy dziecko. Jeżeli jego koleżanki i koledzy wywodzą się z tak zwanych trudnych środowisk i sami mają znikome widoki na powodzenie życiowe, to naszemu potomstwu prawdopodobnie też znacznie trudniej przyjdzie uniknąć biedy.

Analiza danych mówi więc, że istnieją takie miejsca w Ameryce, gdzie dzieci mają większe szanse, by uciec przed ubóstwem. A gdzie panują najlepsze warunki, żeby się schować przed Ponurym Żniwiarzem?[4]

*

Często myślimy o śmierci jako o „wielkim wyrównywaczu", bo przecież nie uniknie jej nikt z nas – ani żebrak, ani król, ani bezdomny, ani Mark Zuckerberg.

Owszem, ludzie zamożni też nie ujdą śmierci, lecz analiza danych pokazuje, że mogą ją opóźnić. W Stanach Zjednoczonych kobiety plasujące się w górnym procencie osób o najwyższych dochodach żyją przeciętnie o dziesięć lat dłużej niż kobiety z przeciwległego krańca skali zarobków; dla mężczyzn różnica ta wynosi piętnaście lat.

A jak kształtują się te dysproporcje w różnych częściach kraju? Czy długość życia wygląda inaczej w zależności od miejsca zamieszkania? Czy jest inna dla biednych i bogatych? Dzięki badawczemu zbliżeniu na wybrane części USA zespół Chetty'ego i tym razem znalazł odpowiedzi.

I co ciekawe, okazało się, że długość życia najbogatszych Amerykanów wcale nie zależy od tego, gdzie mieszkają. Kobieta, która dysponuje nadwyżką gotówki, ma szanse dożyć około osiemdziesięciu dziewięciu lat, a mężczyzna około osiemdziesięciu siedmiu. Ludzie zamożni na całym świecie prowadzą zwykle zdrowszy tryb życia – częściej ćwiczą, lepiej się odżywiają, palą mniej papierosów i rzadziej cierpią na otyłość. Bogatych stać na bieżnię do biegania w domu, ekologiczne awokado i zajęcia z jogi, i mogą je sobie zapewnić w dowolnym zakątku Stanów Zjednoczonych.

Niestety ludzie ubodzy to całkiem inna historia – w ich wypadku długość życia różni się dramatycznie w zależności od miejsca zamieszkania. Ale jeżeli dobrze trafią pod tym względem, mogą przedłużyć sobie żywot doczesny nawet o pięć lat.

Dlaczego w niektórych miejscach ludzie ubodzy żyją znacznie dłużej niż gdzie indziej? I czy jest coś, co łączy miasta, gdzie biedni żyją najdłużej?

Poniżej wymieniam cztery czynniki miejskie, z których tylko jeden ma wyraźny wpływ na długość życia mieszkańców. Spróbujcie zgadnąć, który.

Co sprawia, że ubodzy mieszkańcy miast żyją dłużej?

Większość mieszkańców jest religijna.
Niski poziom zanieczyszczenia powietrza.
Znaczna liczba mieszkańców ma ubezpieczenie zdrowotne.
W mieście zamieszkuje wielu ludzi zamożnych.

Pierwsze trzy czynniki – religijność, zanieczyszczenie środowiska i ubezpieczenie zdrowotne – nie mają wpływu na długość życia ubogich mieszkańców. Natomiast badania Chetty'ego i jego kolegów dowodzą, że ważne jest tutaj to, ilu bogatych ludzi mieszka w danym mieście. Im więcej, tym dłużej żyją ich ubodzy współmieszkańcy. Na przykład biedacy nowojorscy żyją o wiele dłużej niż biedacy w Detroit.

Jak to możliwe, że bogacze przedłużają życie biedakom? Jedną z hipotez w odpowiedzi na to pytanie postawił David Cutler, współautor wspomnianych badań, należący do grona moich doradców. Według niego przyczyną może być tak zwane zaraźliwe zachowanie[5].

Wiele wskazuje bowiem na to, że nasze zachowania bywają „zaraźliwe" – dlatego ludzie ubodzy mieszkający w bliskim sąsiedztwie bogatych często przejmują od nich różne zwyczaje. Niektóre, jak, powiedzmy, używanie pretensjonalnego słownictwa, raczej nie wpłyną na stan ich zdrowia, ale inne, na przykład ćwiczenia fizyczne, z pewnością mogą przynieść dobre skutki. I rzeczywiście – biedacy mieszkający w otoczeniu bogaczy częściej ćwiczą, palą mniej papierosów i rzadziej cierpią na otyłość.

Moje ulubione badanie zespołu Raja dotyczy jednak tego, z jakich powodów jedni ludzie oszukują urząd skarbowy w zeznaniach podatkowych, a inni nie[6]. Wyjaśnienie, na czym to badanie polegało, jest jednak trochę bardziej skomplikowane.

Musimy zacząć od stwierdzenia, że w Stanach Zjednoczonych osoby samozatrudnione i wychowujące jedno dziecko mogą podnieść swoje dochody dzięki finansowemu wsparciu państwa. Jeżeli zeznasz, że miałeś w danym roku dochód w wysokości dokładnie dziewięciu tysięcy dolarów, to dostaniesz od rządu federalnego czek na tysiąc trzysta siedemdziesiąt siedem dolarów – kwotę określaną mianem Earned Income Tax Credit, będącą darowizną, która ma wspomóc płatników o niskich dochodach, odejmowaną od podatku od wynagrodzeń. Jeżeli zgłosisz większe dochody, twój podatek od wynagrodzeń wzrośnie, a jeżeli mniejsze, zmniejszy się też twój Earned Income Tax Credit. A zatem zgłosić dochód w wysokości dziewięciu tysięcy dolarów to jakby trafić w dziesiątkę.

Dlatego nietrudno zgadnąć, że osoby samozatrudnione i mające na utrzymaniu jedno dziecko najczęściej zgłaszają dochód w takiej właśnie wysokości.

Czy ci płatnicy przygotowali sobie odpowiedni plan pracy, żeby zarobić dokładnie taką kwotę? Nie. Losowy audyt (choć przeprowadza się go bardzo rzadko) wykazywał prawie zawsze, że nawet nie byli blisko dziewięciu tysięcy dolarów – zarabiali albo znacznie mniej, albo znacznie więcej.

A to oznacza, że dopuszczali się przestępstw podatkowych, zgłaszając dochody w takiej wysokości, która zapewniała im czek na najwyższą możliwą sumę w prezencie od państwa.

Jak często dochodziło do takich oszustw i kto je najczęściej popełniał? Dzięki badaniom Chetty'ego i jego kolegów okazało się, że wyglądało to bardzo różnie w różnych częściach kraju. Na przykład dochody w wysokości dziewięciu tysięcy dolarów zgłosiło zdumiewająco dużo, bo aż trzydzieści procent samozatrudnionych podatników z jednym dzieckiem w Miami – podczas gdy w Filadelfii było ich tylko dwa procent.

Jak można przewidzieć, kto dokona tego rodzaju machlojki? Czy miejsce zamieszkania ma jakiś związek z wysoką albo niską liczbą kłamców podatkowych? Korelacja częstotliwości oszustw z innymi miejskimi czynnikami demograficznymi wykazuje, że większe

znaczenie mają jedynie dwa z nich: duża liczba osób kwalifikujących się do Earned Income Tax Credit i duża liczba doradców podatkowych.

Na co to wskazuje? Chetty i jego zespół podali wytłumaczenie. Kluczowym bodźcem do tego typu kantów była… informacja, a może raczej niedoinformowanie.

Większość samozatrudnionych podatników mających jedno dziecko na utrzymaniu po prostu nie wiedziała, że magiczna liczba, dzięki której mogą zainkasować przyjemną sumkę od państwa, wynosi dziewięć tysięcy dolarów – ale ponieważ mieszkali w otoczeniu osób, które mogły to wiedzieć, jak choćby ich sąsiedzi czy doradcy podatkowi, szanse, że i oni się o tym dowiedzą, znacząco wzrastały.

Zespół Chetty'ego znalazł zresztą jeszcze więcej poszlak, że ten rodzaj oszustw podatkowych napędzała wiedza. Okazało się między innymi, że ludzie, którzy przeprowadzali się z miasta, gdzie podobne szwindle były rzadkością, do miasta, gdzie występowały często, uczyli się tej sztuczki i szybko zaczynali ją stosować, z czasem rozprzestrzeniła się więc po całych Stanach Zjednoczonych. Oszustwa podatkowe okazują się równie zaraźliwe jak wirusy.

A teraz zastanówcie się chwilę, jak odkrywczy jest wynik tego badania – albowiem wskazuje on, że w prognozie, kto dokona oszustwa podatkowego, a kto nie, najważniejszej roli wcale nie odgrywa to, kto jest uczciwy z natury: liczy się to, kto wie, jak dokonać przekrętu, oraz kto tego nie wie.

Kiedy więc ktoś wam powie, że nigdy nie oszukałby państwa na podatkach, najprawdopodobniej – tak, zgadliście – kłamie. Badania Chetty'ego dowodzą, że wielu z nas oszukiwałoby, gdyby wiedziało jak.

Jeśli chcecie zatem oszukiwać (chociaż w żadnym wypadku do tego nie namawiam), to powinniście zamieszkać w otoczeniu zawodowych doradców podatkowych albo innych macherów, którzy powiedzą wam, co macie robić.

A gdzie należy mieszkać, jeżeli chce się mieć sławne na cały świat dzieci? Filmowe zbliżenie na dane analityczne i ich możliwie największe uszczegółowienie pomoże nam odpowiedzieć i na to pytanie.

Byłem ciekawy, skąd pochodzą Amerykanie odnoszący największe sukcesy życiowe, pewnego dnia postanowiłem więc ściągnąć sobie na twardy dysk całą Wikipedię (w dzisiejszych czasach można już robić takie rzeczy)[7].

Napisałem krótki program i stworzyłem zbiór danych obejmujący ponad sto pięćdziesiąt tysięcy Amerykanów, których redaktorzy Wikipedii uznali za na tyle rozpoznawalnych, żeby umieścić ich na swoich stronach. Zbiór ten zawierał takie informacje, jak kraj i datę urodzenia, zawód oraz płeć. Następnie zmieszałem te dane ze zbieranymi na potrzeby National Center for Health Statistics świadectwami urodzenia z wszystkich hrabstw w Stanach. I obliczyłem, jakie szanse trafienia do Wikipedii mają Amerykanie z poszczególnych hrabstw.

Czy strona w encyklopedii internetowej rzeczywiście jest istotnym wyznacznikiem sukcesu? Nie zawsze. Redaktorzy preferują młodych mężczyzn, co ujmuje trochę obiektywizmu temu zbiorowi danych. Poza tym takie wyróżnienie nie zawsze jest przecież zaszczytne. Na przykład Ted Bundy znalazł się w Wikipedii, bo zamordował co najmniej trzydzieści kobiet. Udało mi się jednak wyeliminować przestępców, nie wypaczając zbytnio ostatecznych rezultatów.

Ograniczyłem się do „baby boomersów" – osób urodzonych między 1946 i 1964 rokiem – ponieważ miały one prawie całe życie na to, aby zyskać sławę. Okazało się, że mniej więcej jedno urodzone w tym czasie w USA dziecko na dwa tysiące pięćdziesiąt osiem osiągnęło wystarczający sukces, by trafić do Wikipedii – z tego trzydzieści procent dzięki dokonaniom na polu sztuki albo show-biznesu, dwadzieścia dziewięć w sporcie, dziewięć w polityce i trzy w nauce albo za sprawą zasług akademickich.

Pierwszym uderzającym faktem, jaki zauważyłem, było ogromne zróżnicowanie geograficzne, jeżeli chodzi o szanse na duży sukces, przynajmniej rozumiany w kategoriach Wikipedii – okazało się bowiem, że szanse te w znacznej mierze zależały od tego, gdzie człowiek się urodził.

Do Wiki trafia mniej więcej jedno na tysiąc dwadzieścioro dziewięcioro dzieci przychodzących na świat w Kalifornii, ale tylko jedno na cztery tysiące czterysta dziewięćdziesięcioro sześcioro w Wirginii Zachodniej. Jeśli spojrzymy na poszczególne hrabstwa, sytuacja stanie się jeszcze jaśniejsza. Swoją stronę w Wikipedii ma mniej więcej jedno na siedemset czterdzieścioro ośmioro dzieci z hrabstwa Suffolk w Massachusetts, w którym leży Boston, są jednak hrabstwa, gdzie szanse na sukces okazały się dwadzieścia razy mniejsze.

Dlaczego niektóre części USA wydają się o wiele lepszym miejscem dla produkcji naszych sław i gwiazd? Przypatrzyłem się uważnie hrabstwom z czołówki i okazuje się, że prawie wszystkie z nich mieszczą się w jednej z dwóch kategorii.

W pierwszej kolejności zaskoczyło mnie to, że w wielu z takich uprzywilejowanych hrabstw leżały spore miasta uniwersyteckie. Prawie za każdym razem, kiedy wysoko na liście widziałem nazwę jakiegoś nieznanego mi wcześniej hrabstwa, jak Washtenaw w Michigan, okazywało się, że jego główną atrakcją jest uniwersytet, w tym wypadku w Ann Arbor. Hrabstwa, których ozdobą są takie miejscowości jak Madison (w stanie Wisconsin), Athens (w stanie Georgia), Columbia (w stanie Missouri), Berkeley (w stanie Kalifornia), Chapel Hill (w stanie Karolina Północna), Gainesville (w stanie Floryda), Lexington (w stanie Kentucky) i Ithaca (w stanie Nowy Jork), plasują się w trzech procentach przewodzących stawce.

Dlaczego tak się dzieje? Odpowiedź może po części kryć się w puli genowej: dzieci profesorów i magistrów są zwykle inteligentne (a w grze o wielki sukces inteligencja może być niezwykle użyteczna). Naprawdę wygląda więc na to, że im więcej osób z wyższym wykształceniem mieszka w danym mieście, tym większe szanse na sukces mają ludzie, którzy się w nim urodzili.

Najprawdopodobniej ważną rolę odgrywa jednak jeszcze jeden czynnik: wczesny kontakt z innowacyjnymi, oryginalnymi pomysłami. Jedną z dziedzin, w której miasta uniwersyteckie produkują swoje znakomitości, jest muzyka. W mieście uniwersyteckim młody chłopak może chodzić na niezwykłe koncerty, słuchać niekonwencjonalnych

rozgłośni radiowych, a nawet odwiedzać niezależne sklepy płytowe. Sprawa nie kończy się jednak na sztuce. W miastach uniwersyteckich kształtuje też swoją osobowość większa, niż można by przypuszczać, liczba znanych później powszechnie ludzi biznesu. Być może wczesna styczność z awangardową sztuką i pomysłami okazuje się pomocna i dla nich.

Sukces miast uniwersyteckich jest niezależny nie tylko od lokalizacji, ale też od koloru skóry ich mieszkańców. Jeśli nie liczyć sportu, Afroamerykanów w Wikipedii jest wyraźnie niewielu, zwłaszcza w biznesie i nauce. Niewątpliwie ma to duży związek z dyskryminacją. Ale znalazłem jedno małe hrabstwo, gdzie w 1950 roku ludność była w osiemdziesięciu czterech procentach czarna, a mimo to ludzie sukcesu wychodzili stamtąd niemal z taką samą częstotliwością jak z hrabstw na szczycie listy.

Z niecałych trzynastu tysięcy „baby boomersów" urodzonych w hrabstwie Macon w Alabamie do Wikipedii trafiło piętnaścioro – czyli jedno na osiemset pięćdziesięcioro dwoje dzieci. I wszystkie z nich są czarne. Czternaście osób pochodziło z Tuskegee, gdzie mieści się uniwersytet o tej samej nazwie, historyczna uczelnia założona przez Bookera T. Washingtona. Są wśród tych osób sędziowie, pisarze i uczeni. Czarnoskóre dziecko z Tuskegee miało zatem takie same szanse, aby osiągnąć sukces w dziedzinie pozasportowej, jak białe, pochodzące z plasujących się najwyżej na liście i w większości zamieszkanych przez białych miast uniwersyteckich.

Drugim elementem, który najbardziej przyczyniał się do sukcesów dzieci urodzonych w danym hrabstwie, były duże metropolie[8]. Ten, kto przyszedł na świat w hrabstwie San Francisco, Los Angeles czy Nowy Jork, miał największe szanse trafić do Wikipedii (potraktowałem pięć nowojorskich hrabstw wspólnie, ponieważ wiele haseł nie uwzględniało dzielnicy, w której urodził się dany człowiek sukcesu).

W obszarze miejskim znajdziemy wiele przykładów udanej kariery i awansu. Żeby zobaczyć, czy warto za młodych lat mieszkać w otoczeniu mistrzów swojego rzemiosła, porównajmy Nowy Jork, Boston i Los Angeles. Z tych trzech miast Nowy Jork produkuje najwięcej dziennikarzy, Boston naukowców, a Los Angeles aktorów.

Pamiętajcie tylko, że mówimy o sławnych ludziach, którzy urodzili się w tych miastach, a nie przenieśli tam z innych części kraju. I nawet jeśli wykluczymy z tego grona osoby, których rodzice osiągnęli wcześniej mistrzostwo w swojej dziedzinie, to Nowy Jork pozostanie miastem dziennikarzy, Boston uczonych, a LA ludzi sceny i ekranu.

Hrabstwa podmiejskie i prowincjonalne, jeżeli nie leżały w nich ważne ośrodki uniwersyteckie, wypadły znacznie słabiej. Moi rodzice, jak wielu „baby boomersów", zamienili zatłoczone ulice na cień drzew – przeprowadzili się z Manhattanu do hrabstwa Bergen w New Jersey – żeby tam wychowywać troje swoich potomków. Potencjalnie rzecz biorąc, był to błąd, w każdym razie jeśli chcieli mieć sławne dzieci, bo te urodzone w Nowym Jorku mają osiemdziesiąt procent więcej szans trafić do Wikipedii niż chłopak z hrabstwa Bergen. To tylko korelacje, ale pokazują, że pod tym względem lepiej jest dorastać w otoczeniu wielkich idei i myśli niż na wielkim podwórku.

Wyniki badań, które tu podaję, mogłyby być jeszcze dokładniejsze, gdybym miał lepsze dane na temat miejsc, w których znani ludzie mieszkali w dzieciństwie, bo przecież wielu z nich dorastało w innych hrabstwach niż te, gdzie przyszli na świat.

Sukces dużych ośrodków miejskich i miast uniwersyteckich jest uderzający, kiedy po prostu patrzy się na materiały źródłowe. Ja jednak postanowiłem zanurzyć się w nich głębiej, by przeprowadzić bardziej wyrafinowaną analizę empiryczną.

I ustaliłem, że istnieje jeszcze jedna zmienna, wyraźnie wskazująca na to, że dana osoba zapewni sobie miejsce w Wikipedii: liczba imigrantów w hrabstwie jej urodzenia. Im większy jest w nim procent mieszkańców urodzonych poza Stanami Zjednoczonymi, tym większy odsetek dzieci z tego hrabstwa osiąga w przyszłości sukces (i co pan na to, panie Trump?). Z dwóch hrabstw o podobnej populacji miejskiej i uniwersyteckiej więcej sławnych Amerykanów ukształtuje to, które zamieszkuje więcej przybyszów z zagranicy. Dlaczego tak się dzieje?

Przede wszystkim należy pamiętać, że sukces dotyczy głównie dzieci imigranckich. Zbadałem dokładnie biografie stu najsławniejszych białych „baby boomersów" na podstawie materiałów zgromadzonych w projekcie Pantheon, realizowanym przez Massachusetts Institute of

Technology i również korzystającym z danych z Wikipedii. Większość tych osób to artyści, a przynajmniej trzynaścioro z nich miało matki urodzone poza granicami USA, w tym Oliver Stone, Sandra Bullock i Julianne Moore. To dla tamtego okresu wynik trzykrotnie lepszy niż średnia krajowa (wiele sław, na przykład Steve Jobs i John Belushi, miało także ojców imigrantów, ale te dane trudniej było porównać ze średnią krajową, bo informacje o ojcach nie zawsze figurują na świadectwie urodzenia).

A jakie zmienne dowodzą, że dziecko nie odniesie sukcesu? Jedna, która wydała mi się więcej niż trochę zaskakująca, to nakłady finansowe przeznaczane na edukację przez poszczególne stany. W stanach o podobnym odsetku osób mieszkających w miastach wydatki na oświatę i kształcenie nie miały wpływu na liczbę pochodzących stamtąd sławnych pisarzy, artystów czy ludzi biznesu.

Ciekawie wypada porównanie moich interpretacji danych z Wikipedii z rezultatami jednego z omawianych wcześniej badań zespołu Chetty'ego. Przypomnijcie sobie, jak Chetty szukał między innymi takich miejsc, które pozwalają mieszkańcom awansować do wyższej klasy średniej. Ja sprawdzałem, gdzie ludzie mogą osiągnąć rozgłos. Rezultaty okazały się uderzająco rozbieżne.

Wysokie nakłady na edukację ułatwiają dzieciom wejście do wyższej klasy średniej, ale nie bardzo pomagają im zostać liczącym się pisarzem, artystą czy biznesmenem. Wielu sławnych ludzi nienawidziło szkoły, a niektórzy w ogóle ją rzucili.

Chetty ustalił, że Nowy Jork nie jest szczególnie dobrym miejscem do wychowania dziecka, jeśli chcecie mu zapewnić awans w szeregi wyższej klasy średniej – natomiast ja wykazałem, że jest to świetne miejsce, jeżeli pragniecie dać dziecku szansę zdobycia sławy.

Kiedy spoglądamy na czynniki przyczyniające się do sukcesu, poważne różnice pomiędzy hrabstwami zaczynają wydawać się logiczne, bo wiele z tych hrabstw posiada większość najważniejszych cech potrzebnych do tego, żeby zapewnić nam życiowy sukces. Wróćmy raz jeszcze do Bostonu i okolic. Znajduje się tam wiele wyższych uczelni, więc miasto aż kipi od najróżniejszych innowacyjnych pomysłów i idei. Jest to rejon zurbanizowany, gdzie mnóstwo naprawdę

wybitnych osób daje przykład młodzieży, jak można zrobić karierę, i który przyciąga wielu imigrantów, ich dzieci zaś zyskują motywację, aby iść za przykładem ludzi sukcesu.

A jeżeli w danej okolicy brakuje tych wartości? Czy to znaczy, że będzie ona produkować mniejszą liczbę supergwiazd? Niekoniecznie. Istnieje inna ścieżka: wąska specjalizacja. Jej dobrym przykładem jest wiejskie hrabstwo Roseau w stanie Minnesota, gdzie mieszka niewielu obcokrajowców i nie ma żadnych ważniejszych uczelni wyższych – ale i tak jedna na siedemset czterdzieści osób tam urodzonych znalazła się w Wikipedii. Ich sekret? Wszyscy są zawodowymi hokeistami, którym w osiągnięciu sławy i pieniędzy niewątpliwie pomogły miejscowe, słynne na cały świat licealne i młodzieżowe programy szkoleniowe.

Czy zatem zakładając, że nie jesteś specjalnie zainteresowany wychowaniem gwiazdy hokeja na lodzie, oznacza to, że musisz się przeprowadzić do Bostonu albo Tuskegee, by stworzyć swoim dzieciom jak najlepsze warunki rozwoju? Z pewnością w niczym by im to nie zaszkodziło. Ale możemy się z tych przykładów nauczyć czegoś ważniejszego. Ekonomiści i socjologowie skupiają się zwykle na tym, jak uniknąć zła w postaci ubóstwa i przestępczości. Zadaniem naprawdę wielkiego narodu nie jest jednak tylko jak największe ograniczenie liczby obywateli, którzy zostają w tyle, lecz pomoc jak największej liczbie potencjalnych sław. Być może takie badania, biorące pod lupę miejsca, gdzie urodziły się setki tysięcy najsławniejszych Amerykanów, pozwolą nam przyjąć wstępną strategię w tym względzie: powinniśmy zatem między innymi promować imigrację, finansować wyższe uczelnie i wspierać artystów i sztukę.

Zwykle analizuję dane dotyczące głównie Stanów Zjednoczonych, więc kiedy myślę o braniu pod lupę różnych miejsc na mapie naszego globu, przychodzą mi do głowy przede wszystkim nasze, amerykańskie miasta i miasteczka. To dlatego przyglądam się takim miejscom jak hrabstwo Macon w Alabamie czy Roseau w Minnesocie. Ale kolejnym wielkim i wciąż rosnącym plusem danych internetowych jest to, że dzięki nim można łatwo zbierać informacje z całego świata i zobaczyć, jak poszczególne kraje różnią się między sobą. Analitycy danych mają

w ten sposób okazję, żeby po cichutku, na paluszkach, wkroczyć na teren antropologii.

Niedawno zająłem się na przykład takim tematem: jak ludzie w różnych krajach podchodzą do ciąży? Przyjrzałem się wpisom w Google'u dokonywanym przez panie będące przy nadziei. I pierwszą rzeczą, jaką zauważyłem, była niezwykła zgodność co do fizycznych symptomów ciąży, na jakie się uskarżały.

Sprawdziłem, jak licznie wyszukiwały różne dolegliwości w połączeniu z określeniem „w ciąży". Na przykład jak często ciąża pojawia się w kontekście torsji, bólu pleców albo zatwardzenia. Wyszukiwania w Kanadzie okazały się bardzo zbliżone do amerykańskich; podobnie w Wielkiej Brytanii, Australii czy Indiach.

Ciężarne jak świat długi i szeroki najwyraźniej mają także apetyt na podobne produkty. W Stanach Zjednoczonych najczęstszym wpisem w tej kategorii jest „apetyt na lody w czasie ciąży". Dalej idą sól, słodycze, owoce i pikantne potrawy. W Australii jest dość podobnie: na liście znajdują się sól, słodycze, czekolada, lody i owoce. A w Indiach? Też podobnie: pikantne potrawy, słodycze, czekolada, sól i lody. Pięć pierwszych miejsc wygląda prawie tak samo we wszystkich krajach, którym się przyglądałem.

Wstępne badania wskazują, że nigdzie na świecie nie znaleziono diety ani warunków środowiskowych, które miałyby wyraźny wpływ na zmianę fizycznego doświadczania ciąży.

Ale w różnych krajach zdecydowanie inaczej o niej myślimy.

Zacznijmy od pytania, co kobiety w ciąży mogą bezpiecznie robić. W USA najczęściej chcemy wiedzieć, czy mogą jeść krewetki, pić wino, kawę albo zażywać tylenol.

W tej sferze inne kraje nie mają wiele wspólnego ani ze Stanami Zjednoczonymi, ani ze sobą. To, czy kobietom w ciąży wolno pić wino, nie należy do pierwszej dziesiątki pytań w Kanadzie, Australii czy Wielkiej Brytanii. W Australii panie martwią się głównie tym, czy mogą w ciąży spożywać produkty mleczne, zwłaszcza serek twarogowy. W Nigerii, gdzie tylko trzydzieści procent populacji korzysta

z Internetu, pierwsze pytanie brzmi, czy kobiety w ciąży mogą pić zimną wodę.

A czy obawy pytających są zasadne? To zależy. Istnieje sporo poszlak wskazujących na to, że kobiety w ciąży są bardziej narażone na listeriozę, jeżeli jedzą sery z niepasteryzowanego mleka. Wiadomo też, że niekorzystnie na stan zdrowia dziecka wpływa nadmierne spożycie alkoholu. W niektórych częściach świata ludzie uważają, że picie zimnej wody może wywołać u noworodka zapalenie płuc, nie znalazłem jednak żadnego medycznego potwierdzenia tych przypuszczeń.

Poważne różnice międzynarodowe w pytaniach dotyczących ciąży wynikają najprawdopodobniej z zalewu informacji, płynących w każdym kraju z różnych źródeł: z rzetelnych i mniej rzetelnych badań naukowych, z plotek albo rozmów z sąsiadami, dlatego kobiety nie za dobrze wiedzą, na czym powinny się skupić – i o co pytać Google'a.

Kiedy spojrzymy na najczęstsze wpisy ciążowe, dostrzeżemy też jeszcze jedną oczywistą różnicę: w Stanach Zjednoczonych, Australii i Kanadzie najpowszechniejszym pytaniem jest „jak zapobiec rozstępom w czasie ciąży". Ale w Ghanie, Indiach i Nigerii metody zapobiegania rozstępom nie znajdują się nawet w pierwszej piątce. W tych krajach panie martwią się bardziej o to, jak uprawiać seks i w jakiej pozycji spać.

Pięć kolejnych najczęstszych wyszukiwań typu „Jak... w czasie ciąży"

USA	Indie	Australia	Wielka Brytania	Nigeria	RPA
jak zapobiegać rozstępom	w jakiej pozycji spać	jak zapobiegać rozstępom	jak stracić na wadze	jak uprawiać seks	jak uprawiać seks
jak stracić na wadze	jak się kochać	jak stracić na wadze	jak zapobiegać rozstępom	jak stracić na wadze	jak stracić na wadze
jak uprawiać seks	jak uprawiać seks	jak uniknąć rozstępów	jak uniknąć rozstępów	jak się kochać	jak zapobiegać

					rozstępom
jak uniknąć rozstępów	jak uprawiać miłość	w jakiej pozycji spać	w jakiej pozycji spać	jak zachować zdrowie	w jakiej pozycji spać
jak zachować sprawność fizyczną	jak dbać o siebie	jak uprawiać seks	jak uprawiać seks	co zrobić, żeby przestać wymiotować	co zrobić, żeby przestać wymiotować

Pięć kolejnych najczęstszych wyszukiwań zaczynających się od „Czy kobieta w ciąży może...”

USA	jeść krewetki	pić wino	pić kawę	zażywać tylenol	jeść sushi
Wielka Brytania	jeść krewetki	jeść wędzonego łososia	jeść sernik	jeść mozzarellę	jeść majonez
Australia	jeść serek twarogowy	jeść krewetki	jeść boczek	jeść śmietanę	jeść ser feta
Nigeria	pić zimną wodę	pić wino	pić kawę	uprawiać seks	jeść liście, owoce i przetwory z moringi (miejscowe drzewo)
Singapur	pić zieloną herbatę	jeść lody	jeść duriany	pić kawę	jeść ananasy
Hiszpania	jeść pasztet	jeść szynkę	zażywać paracetamol	jeść tuńczyka	opalać się
Niemcy	latać samolotem	jeść salami	chodzić do sauny	jeść miód	jeść mozzarellę
Brazylia	farbować włosy	zażywać dypiron (środek przeciwbólowy)	zażywać paracetamol	jeździć na rowerze	latać samolotem

Nie ulega wątpliwości, że możemy się bardzo wiele nauczyć, przyglądając się różnym aspektom zdrowia i kultury w różnych zakątkach świata. Moja wstępna analiza wskazuje jednak, że gdy chodzi o możliwość wykroczenia poza swoją biologiczną naturę, człowiek jest jeszcze bardziej bezradny, niż sądziliśmy. Tyle tylko, że zaskakująco różnie interpretujemy ten fakt w różnych krajach.

JAK SPĘDZAMY KOLEJNE MINUTY I GODZINY SWOJEGO ŻYCIA

„Przygody młodzieńca, zainteresowanego głównie gwałtem, skrajną przemocą i Beethovenem".

W ten sposób reklamowano kontrowersyjną *Mechaniczną pomarańczę* Stanleya Kubricka. Młody bohater filmu, Alex DeLarge, dopuszczał się szokujących aktów przemocy, zachowując przy tym mrożącą krew w żyłach obojętność wobec swoich ofiar. W jednej z najsłynniejszych scen gwałci kobietę, rycząc jednocześnie *Singin' in the Rain*.

Niemal natychmiast po premierze pojawiły się doniesienia o przypadkach naśladowania bohatera. Rzeczywiście, grupa mężczyzn zgwałciła siedemnastoletnią dziewczynę, śpiewając tę samą piosenkę. Film został zakazany w wielu krajach europejskich, a niektóre wstrząsające sceny wycięto w wersji pokazywanej w Stanach Zjednoczonych.

Można podać zresztą wiele przykładów naśladowania sztuki przez prawdziwe życie, bo ludzie czasem zachowują się jak urzeczeni tym, co oglądają na ekranie[9]. Po premierze gangsterskiego filmu *Kolory* doszło do potężnej strzelaniny, a po pokazie *New Jack City* wybuchły zamieszki.

Jednym z najbardziej niepokojących przejawów podobnych zachowań były wydarzenia, które rozegrały się cztery dni po premierowym seansie *Pociągu z forsą*, gdy dwóch facetów na stacji metra w Brooklynie oblało budkę biletera skroplonym gazem do zapalniczek i podpaliło ją, niemal dosłownie powtarzając scenę z filmu. Była tylko

jedna różnica: na ekranie kasjerowi udało się uciec, ale w realnym świecie spłonął żywcem.

Przeprowadzono także pewne eksperymenty psychologiczne świadczące o tym, że po obejrzeniu filmu pełnego przemocy ludzie odczuwają więcej złości i wrogości do świata, nawet jeśli nie próbują naśladować tego, co widzieli[10].

Można by więc pomyśleć, że pełne agresji filmy mogą wywoływać agresywne zachowania u widzów. Ale czy skutki tych zachowań naprawdę są tak znaczące? Czy mówimy o jednym albo dwóch zabójstwach na dziesięć lat, czy o setkach zabitych każdego roku? Doniesienia prasowe ani eksperymenty psychologiczne nie dadzą nam odpowiedzi na to pytanie.

By się przekonać, czy i w tej kwestii może nam pomóc big data, dwaj ekonomiści, Gordon Dahl i Stefano DellaVigna, połączyli trzy duże zbiory danych z lat 1995–2004: dane o przestępczości z FBI, obejmujące rejestrowane z godziny na godzinę wszelkie poważniejsze naruszenia prawa, liczbę widzów w kinach oraz skalę przemocy we wszystkich wyświetlanych w tym czasie filmach według kids-in-mind.com.

Dahl i DellaVigna korzystali z kompletnych informacji – uwzględnili absolutnie wszystkie filmy i przestępstwa popełniane w każdej godzinie we wszystkich miastach Stanów Zjednoczonych. Szczegółowość tych danych miała okazać się ważna.

Kluczowe dla badań Dahla i DellaVigni było to, że w niektóre weekendy najpopularniejszy film wprost kipiał przemocą, jak *Hannibal* albo *Świt żywych trupów*, ale kiedy indziej największym wzięciem cieszyły się obrazy zupełnie pozbawione aktów agresji, jak *Uciekająca panna młoda* czy *Toy Story*.

Nasi ekonomiści widzieli więc doskonale, ile morderstw, gwałtów i napadów zdarzało się w weekendy, kiedy wyświetlano premierowe filmy akcji, i mogli porównać te liczby z liczbą morderstw, gwałtów i napadów, do których dochodziło w weekendy, gdy odbywały się premiery popularnych, lecz spokojnych filmów.

I co się okazało? Czy po premierze agresywnego filmu przestępczość rosła, jak wskazywałyby na to niektóre eksperymenty psychologiczne?

A może pozostawała na tym samym poziomie?

Otóż badania Dahla i DellaVigni dowiodły, że w weekendy premier filmów akcji przestępczość spadała[11].

Nie mylicie się. W dni, kiedy miliony Amerykanów oglądały sceny zabijania, przestępczość była niższa – i to w poważnym stopniu.

Kiedy uzyskujesz taki dziwny i nieoczekiwany wynik, twoją pierwszą myślą jest to, że popełniłeś jakiś błąd. Obaj uczeni starannie jeszcze raz przejrzeli zakodowane przez siebie dane. Żadnych błędów. Druga myśl jest taka, że istnieje jeszcze jedna, nieznana dotąd zmienna, która wyjaśni wyniki. Dahl i DellaVigna sprawdzili więc, czy może wpływa na nie, powiedzmy, pora roku. Okazało się, że nie. Potem zebrali dane o pogodzie, sądząc, że może to ona jakoś kształtuje rezultat. Też nie.

„Sprawdziliśmy wszystkie swoje przypuszczenia, działania i analizy – powiedział Dahl. – Nie znaleźliśmy jednak żadnych uchybień".

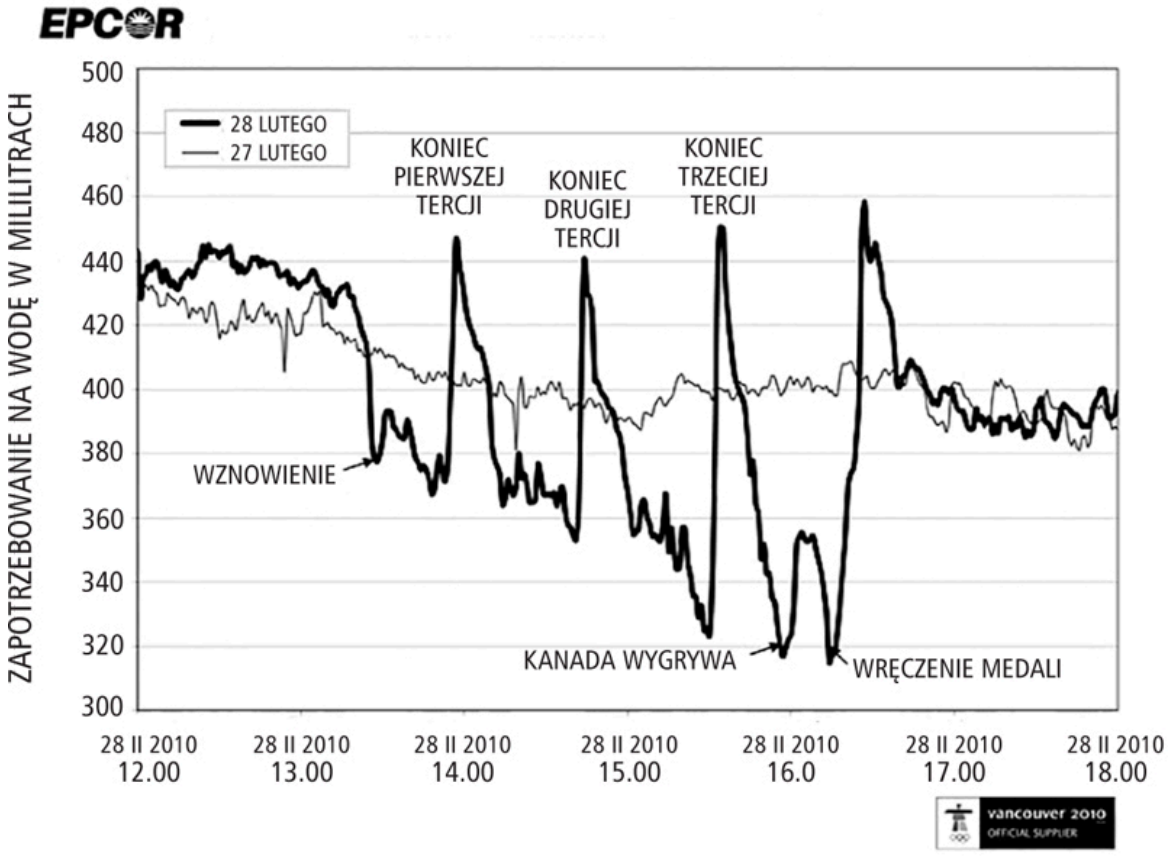
Na przekór doniesieniom prasowym i naukowym doświadczeniom psychologicznym pokaz filmu pełnego przemocy, choć brzmi to bardzo dziwnie, w jakiś sposób powodował znaczny spadek przestępczości. Jak to jest możliwe?

Uczeni mogli odpowiedzieć na powyższe pytanie dzięki temu, że zaprzęgli big data do jeszcze bardziej szczegółowych badań. Dane sondażowe dostarczały zwykle informacji corocznych, w najlepszym wypadku comiesięcznych. Mając ogromne szczęście, można było uzyskać statystyki z pojedynczego weekendu. Ale odkąd zaczęliśmy coraz częściej korzystać z naprawdę wyczerpujących baz informacji zamiast z sondaży, opartych tylko na niewielkiej grupce respondentów, możemy zobaczyć to, co dzieje się z godziny na godzinę albo nawet z minuty na minutę. I dzięki temu dowiadujemy się więcej o naszym behawiorze.

Zdumiewające, a czasami zabawne, bywają fluktuacje temporalne. Na przykład EPCOR, przedsiębiorstwo użyteczności publicznej w kanadyjskim mieście Edmonton, przedstawiło kiedyś dane pokazujące rejestrowane minuta po minucie zapotrzebowanie na wodę w czasie odbywającego się na igrzyskach olimpijskich w 2010 roku meczu o złoty medal w hokeju na lodzie pomiędzy Stanami

Zjednoczonymi i Kanadą. Oglądało go, jak się szacuje, osiemdziesiąt procent Kanadyjczyków. Dane mówią nam, że tuż po zakończeniu każdej tercji raptownie wzrastało zużycie wody w mieście: najwyraźniej mieszkańcy spuszczali ją jednocześnie w większości toalet w Edmonton.

Zużycie wody w Edmonton w czasie meczu hokejowego o złoty medal



Wyszukiwania w Google'u też można badać z dokładnością do jednej minuty, ujawniając przy tym ciekawe prawidłowości[12]. Na przykład wyszukiwanie dostępu do „niezablokowanych gier" skacze w górę o ósmej rano w dni robocze i pozostaje na wysokim poziomie do trzeciej po południu, zapewne dlatego, że szkoły usiłują blokować dostęp do internetowych gier komputerowych w czasie lekcji, a jednocześnie nie

chcą zabraniać uczniom przynoszenia do szkoły telefonów komórkowych.

Wyszukiwania typu „pogoda", „modlitwa" czy „wiadomości" są najczęstsze przed piątą trzydzieści rano, co świadczy o tym, że większość ludzi wstaje znacznie wcześniej niż ja. Wyszukiwania typu „samobójstwo" zdarzają się najczęściej trzydzieści sześć minut po północy, najrzadziej natomiast około dziewiątej rano, co z kolei dowodzi, że większa część ludzkości czuje się o poranku znacznie lepiej niż autor tych słów.

Z analizy danych wynika również, że ważne pytania zadajemy w sieci najchętniej pomiędzy drugą i czwartą rano: „Jakie jest znaczenie świadomości?", „Czy istnieje wolna wola?", „Czy istnieje życie na innych planetach?". Popularność tych pytań o tak późnej porze jest chyba częściowo związana z paleniem marihuany – bo o to, jak skręcić dżointa, pytamy z kolei najczęściej pomiędzy pierwszą i drugą po północy.

Dahl i DellaVigna mogli więc zobaczyć, jak wygląda i zmienia się przestępczość z godziny na godzinę w filmowe weekendy. Stwierdzili, że jej spadek w stosunku do innych dni niż te, kiedy pokazywano filmy pełne agresji, zaczynał się wczesnym wieczorem – czyli jeszcze zanim na ekranie kinowym zaczynały się pojawiać sceny przemocy, gdy widzowie dopiero wchodzili na seans.

Domyślacie się dlaczego? Zastanówcie się najpierw, kto najchętniej wybiera się do kina na brutalny film; zwykle są to młodzi, a przy tym także agresywni ludzie.

A teraz zastanówcie się, gdzie najczęściej popełnia się przestępstwa. W kinie stosunkowo rzadko, chociaż są wyjątki, jak na przykład przeprowadzona z premedytacją strzelanina w kinie w Kolorado. Na ogół jednak chodzimy do kina bez broni i zachowujemy ciszę na sali.

Zaproponujcie młodym, agresywnym mężczyznom obejrzenie *Hannibala*, a pójdą na seans. Zaproponujcie im zamiast tego tytułu *Uciekającą pannę młodą*, a zrezygnują z filmu i pójdą gdzieś indziej, być może do baru, klubu albo na bilard, słowem: do miejsc, gdzie częściej dochodzi do przestępstw z użyciem przemocy.

Czyli filmy akcji zatrzymują osoby potencjalnie agresywne w salach kinowych, dzięki czemu zmniejsza się zagrożenie na ulicach.

Zagadka rozwiązana, prawda? Niezupełnie, bo z analizy wynikało jeszcze coś dziwnego: przestępczość zaczynała spadać tuż przed seansem, lecz nie wzrastała zaraz po jego zakończeniu i zamknięciu kin. Przeciwnie, jej wskaźnik pozostawał niski jeszcze długo w noc, od północy do szóstej rano.

Ale jeżeli przestępczość spadała, kiedy młodzi ludzie wciąż przebywali w kinie, to czy nie powinna wzrosnąć, gdy już z niego wyszli i nie byli zajęci oglądaniem? Przecież dopiero co widzieli pełen przemocy film, powinni więc, jak powiadają wyniki eksperymentów psychologicznych, stawać się gniewniejsi i bardziej agresywni.

Czy przychodzą wam do głowy jakieś wyjaśnienia, dlaczego liczba czynów karalnych po zakończeniu seansu wciąż nie wzrastała? Po długim namyśle autorzy – nawiasem mówiąc, eksperci od przestępczości – doznali nowego olśnienia. Zdawali sobie sprawę, że jednym z głównych czynników sprzyjających wzrostowi bandytyzmu jest alkohol[13], a sami chodzili na tyle często do kin, by wiedzieć, że w Stanach Zjednoczonych właściwie nigdzie nie sprzedaje się w nich mocnych trunków. Mało tego, stwierdzili, że w godzinach nocnych po zakończeniu seansów kinowych przestępczość związana ze spożyciem napojów wyskokowych spadała nader gwałtownie.

Oczywiście badania naszej dwójki uczonych miały ograniczony zasięg. Nie mogli na przykład sprawdzić swoich wyników z punktu widzenia efektów długoterminowych, mierzonych, powiedzmy, w miesiącach – nie mogli zatem zobaczyć, jak długo spadek wskaźnika przestępczości może się utrzymać. Dlatego wcale nie jest wykluczone, że częsta konsumpcja pełnych przemocy filmów ostatecznie rodzi jednak przemoc w realnym świecie. Ale analiza Dahla i DellaVigni przedstawia w innym świetle bezpośrednie oddziaływanie na naszą psychikę krwawych filmów akcji, które stanowiły główny element wspomnianych doświadczeń psychologicznych. Być może ekranowa brutalność oddziałuje na niektórych ludzi i napełnia ich nadzwyczajną złością i agresją. Czy wiecie jednak, co bez żadnych wątpliwości pcha

ludzi na drogę przemocy? Towarzystwo potencjalnie agresywnych kompanów i alkohol[15*].

Dzisiaj to rozumiemy, ale nie rozumieliśmy, zanim Dahl i DellaVigna nie zaczęli analizować całej masy danych na ten temat[15].

Jest jeszcze jedna ważna rzecz, która staje się jasna, kiedy patrzymy na jakieś zjawisko z bliska: świat okazuje się skomplikowany, a działania, które podejmujemy dzisiaj, mogą kiedyś mieć dalekosiężne skutki, w większości niezamierzone. Myśli rozprzestrzeniają się, czasami powoli, a czasem wykładniczo, jak wirusy, ludzie zaś reagują na bodźce nieprzewidywalnie.

Takich relacji i związków, wzrostów i spadków nie da się wyśledzić dzięki małym sondażom czy tradycyjnym metodom gromadzenia informacji. Świat jest po prostu strukturą zbyt złożoną i bogatą, aby do jego poznania mógł wystarczyć niewielki zbiór danych.

NASZE SOBOWTÓRY

W czerwcu 2009 roku wyglądało na to, że David „Big Papi" Ortiz jest skończony. Ale wcześniej przez pięć lat Boston kochał się w swoim urodzonym na Dominikanie pałkarzu o przyjaznym uśmiechu i z przerwą między zębami.

Wystąpił w pięciu kolejnych meczach All Stars, zdobył nagrodę dla najbardziej wartościowego gracza i pomógł drużynie z Bostonu zakończyć trwającą osiemdziesiąt sześć lat posuchę, czyli zwyciężyć w World Series. Ale w sezonie 2008, gdy skończył trzydzieści dwa lata, zaczął grać coraz słabiej. Jego średnia punktowa jako pałkarza spadła o sześćdziesiąt osiem punktów, liczba zdobytych baz o siedemdziesiąt osiem, a procentowe wyniki uderzeń double, triple i home runów o sto czternaście. Na początku sezonu 2009 zrobiło się jeszcze gorzej.

Oto w jaki sposób pisał o nim w tamtym czasie Bill Simmons[16], dziennikarz sportowy i zagorzały kibic zespołu Boston Red Sox: „Stało się jasne, że David Ortiz nie jest już mistrzem bejsbolu… Pałkarze-mięśniacy są niczym gwiazdy porno, zapaśnicy, centrzy z NBA i młode

żony podstarzałych facetów: w pewnym momencie odchodzą". Wielcy kibice sportowi wierzą swoim oczom, a oczy Simmonsa mówiły, że Ortiz jest skończony. Dziennikarz przewidywał nawet, że „Papi" wyląduje na ławce rezerwowych albo że Boston rozwiąże z nim kontrakt.

Czy był to naprawdę koniec Ortiza? Powiedzmy, że w roku 2009 jesteś generalnym menedżerem Bostonu: wywalasz go? Drugie pytanie natury ogólniejszej: skąd możemy wiedzieć, w jakiej formie będzie bejsbolista za jakiś czas?[17] I pytanie trzecie, jeszcze bardziej ogólne: w jaki sposób należy korzystać z big data, żeby przewidzieć, co i jak ludzie będą robić w przyszłości?

Oto, co może was zaprowadzić daleko w analizie danych: spójrzcie na robotę specjalistów od sabermetrii (sabermetria to analiza bejsbolu) i spróbujcie zastosować ich metody na innych polach. Bejsbol był jedną z pierwszych dziedzin, w których dysponowaliśmy pełnymi zbiorami danych na niemal każdy temat związany z tym sportem, mając także istną armię inteligentnych ludzi, gotowych poświęcić całe życie na to, żeby na ich podstawie dowiedzieć się czegoś nowego o ulubionej grze Amerykanów. W dzisiejszych czasach metod sabermetrycznych używa się już albo będzie używać właściwie wszędzie. Bejsbol był pierwszy, inne dziedziny i dyscypliny idą w jego ślady. Sabermetria zawojowuje świat.

Najprostszy sposób, żeby przewidzieć przyszłość bejsbolisty, to założyć, że będzie dalej grał na takim samym poziomie, na jakim gra aktualnie. Jeśli przez ostatnie półtora roku miał kłopoty z formą, to można przypuszczać, że będzie je miał też przez następne osiemnaście miesięcy.

Zgodnie z tą metodą Boston powinien był pozbyć się Ortiza.

Można jednak wykorzystać również inne ważne informacje. W latach osiemdziesiątych Bill James, uznawany na ogół za twórcę sabermetrii, podkreślał znaczenie wieku zawodnika. Bejsboliści, jak twierdził, osiągają szczyt swoich możliwości wcześnie – około dwudziestego siódmego roku życia. A drużyny często nie zwracały uwagi na to, jak bardzo gracze z wiekiem tracą na wartości. I przepłacały za starzejących się zawodników.

Zgodnie z tą bardziej zaawansowaną metodologią Boston zdecydowanie powinien był pozbyć się Ortiza.

Czynnik wieku może być jednak mylący. Nie wszyscy gracze podążają przez życie taką samą ścieżką. Jedni osiągają najlepszą dyspozycję w wieku dwudziestu trzech lat, inni trzydziestu dwóch. Zawodnicy niscy mogą starzeć się inaczej niż wysocy, a grubsi inaczej niż chudzi. Statystycy bejsbolowi dowiedli, że różne typy graczy różnie się starzeją. Ta diagnoza była dla Ortiza jeszcze mniej korzystna, bo pałkarze-mięśniacy zazwyczaj rzeczywiście wcześnie dochodzą do szczytu formy i krótko po trzydziestce tracą swoją wartość[18].

Gdyby zatem Boston rozważył także niedawną przeszłość, wiek, wzrost i wagę Ortiza, to absolutnie i bez wątpienia powinien był się go pozbyć.

Niemniej w 2003 roku statystyk Nate Silver wprowadził nowy model analityczny – PECOTA – służący do prognozowania wyników zawodnika. I ten sposób okazał się najlepszy – a zarazem najbardziej cool. Silver szukał bowiem „sobowtórów" poszczególnych graczy. Wygląda to tak: stwórz bazę danych wszystkich bejsbolistów Major League, licząc od początku jej powstania, czyli złożoną z ponad osiemnastu tysięcy zawodników. Wprowadź do niej wszystko, co o nich wiesz: wzrost, wiek, pozycję na boisku, liczbę home runów, średnią uderzeń, walków i strikeoutów w każdym sezonie. Potem znajdź dwudziestu graczy, którzy najbardziej przypominają Ortiza – tych, którzy grali tak jak on, kiedy miał lat dwadzieścia cztery, dwadzieścia pięć, dwadzieścia sześć, dwadzieścia siedem, dwadzieścia osiem, dwadzieścia dziewięć, trzydzieści, trzydzieści jeden, trzydzieści dwa i trzydzieści trzy. Inaczej mówiąc, znajdź jego „sobowtórów", a potem sprawdź, jak przebiegały ich kariery.

Szukanie sobowtóra to kolejny przykład uszczegółowienia analizy. Przyglądamy się małemu podzbiorowi ludzi najbardziej przypominających daną osobę – i jak zawsze, widać ich tym lepiej, im więcej mamy danych. Tu okazało się, że „sobowtóry" Ortiza – a znaleźli się wśród nich między innymi Jorge Posada i Jim Thome – prognozowały dla niego zupełnie inną przyszłość[19]. Obaj zaczęli karierę dość niemrawo; dobiegając trzydziestki, zanotowali genialne

występy i osiągnęli klasę światową, a zaraz po przekroczeniu trzydziestego roku życia zaczęli mieć kłopoty.

Silver spróbował przewidzieć, co stanie się z Ortizem, na podstawie tego, jak potoczyła się późniejsza kariera Posady i Thome'a – a ci dwaj zawodnicy wrócili w końcu na szczyt. Simmons zapewne ma słuszność co do młodych żon podstarzałych facetów: w pewnym momencie odchodzą. Lecz „sobowtóry" Ortiza odeszły, a potem wróciły.

Ten sposób – „na sobowtóra" – który okazał się najlepszą w historii metodą prognozowania przyszłych osiągnięć bejsbolistów, wskazywał, że Boston powinien zachować cierpliwość wobec swojego starzejącego się pałkarza. W 2010 roku jego średnia wzrosła do dwustu siedemdziesięciu punktów. Miał na koncie trzydzieści dwa home runy i trafił do drużyny All Stars, do której wybierano go przez następne cztery lata. W 2013 roku, wchodząc na boisko tradycyjnie jako trzeci pałkarz, uzyskał sześćset osiemdziesiąt osiem punktów, a Boston wygrał z St. Louis – w meczach cztery do dwóch – i zwyciężył w World Series, Ortiza zaś ogłoszono najbardziej wartościowym graczem tamtych zawodów[16*].

Gdy tylko skończyłem czytać studium Silvera o tym, jak prognozować przyszłość bejsbolistów, natychmiast zacząłem się zastanawiać, czy i ja nie mam swojego sobowtóra.

Metoda „na sobowtóra" jest bowiem obiecująca na wielu polach, nie tylko w sporcie. Czy uda mi się znaleźć kogoś, kto w znacznym stopniu podziela moje zainteresowania? Może gdybym znalazł kogoś bardzo podobnego do siebie, moglibyśmy się spotykać. Może taki facet znałby restauracje, które przypadłyby mi do gustu. I może mógłby mi pokazać coś interesującego, o czym wcześniej nie miałem żadnego pojęcia.

Szukając sobowtórów, skupiamy się na pojedynczych osobach, a nawet na ich poszczególnych cechach. Obraz jest, jak zawsze, tym ostrzejszy, im więcej mamy danych. Przypuśćmy, że szukałem swojego sobowtóra w zbiorze około tysiąca osób. Mógłbym wśród nich znaleźć kogoś, kto lubi książki popularno-naukowe o fizyce. Ale by znaleźć kogoś, kto jest rzeczywiście do mnie podobny pod wieloma względami, musiałbym prowadzić poszukiwania w zbiorze złożonym z setek milionów ludzi.

Pewnego dnia udałem się na polowanie na sobowtóra do mediów społecznościowych. Korzystając z wielkiego korpusu profili na Twitterze, na całej naszej planecie szukałem ludzi podzielających moje podstawowe zainteresowania.

Na pewno można dużo powiedzieć o moich pasjach na podstawie tego, kogo obserwuję na swoim koncie twitterowym. W sumie jest to blisko dwieście pięćdziesiąt osób, podzielających moją namiętność do sportu, polityki, komedii, nauk ścisłych i melancholijnych żydowskich wokalistów folkowych.

Czy zatem jest na świecie ktoś, kto obserwuje wszystkie te same dwieście pięćdziesiąt kont co ja, jakiś mój twitterowy bliźniak? Oczywiście, że nie. Sobowtóry nie są takie same, jak my, tylko do nas podobne. Nie ma zresztą nikogo, kto obserwowałby nawet dwieście tych kont co ja. Ani nawet sto pięćdziesiąt.

W końcu jednak znalazłem konto, które obserwowało – co jest zdumiewające – aż sto profili obserwowanych również przeze mnie: konto rozgłośni radiowej Country Music Radio Today. Co takiego? Okazało się, że Country Music Radio Today było (bo dzisiaj już nie istnieje) botem, obserwującym siedemset pięćdziesiąt profili twitterowych w nadziei, że ich użytkownicy będą obserwować jego.

Chodziłem kiedyś z dziewczyną, która, jak sądzę, miałaby ubaw z tego odkrycia, bo powiedziała mi pewnego razu, że bardziej przypominam robota niż człowieka.

Ale żarty na bok, gdyż owo wstępne odkrycie, że mój sobowtór to bot, który obserwował aż siedemset pięćdziesiąt przypadkowych profili, dowodzi w istocie czegoś ważnego, jeśli idzie o poszukiwania sobowtórów – bo żeby takie poszukiwania zakończyły się prawdziwym powodzeniem, nie można szukać tylko kogoś, kto po prostu lubi to samo co wy: należy bowiem znaleźć kogoś, kto przy tym nie lubi też tych samych rzeczy.

Moje zainteresowania widać nie tylko po kontach, które obserwuję, ale też po kontach, których nie obserwuję. Interesuję się sportem, polityką, komedią i naukami ścisłymi, ale nie kuchnią, modą czy teatrem. Z obserwowanych przeze mnie kont wynika, że lubię Berniego Sandersa, choć nie Elizabeth Warren, lubię Sarah Silverman, ale nie

Amy Schumer, „New Yorkera", ale nie „Atlantic", moich znajomych Noaha Poppa, Emily Sands i Josha Gottlieba, ale nie Sama Ashera (sorki, Sam, ale twoje wpisy na Twitterze to czyste nudziarstwo).

A więc kto z dwustu milionów ludzi na Twitterze ma profil najbardziej podobny do mojego? Okazuje się, że moim sobowtórem jest jeden z autorów „Voxu", Dylan Matthews. Trochę mnie to rozczarowało, bo chciałem nieco poszerzyć swoje horyzonty medialne, a Matthewsa obserwuję już na Twitterze i Facebooku i nałogowo czytam jego posty w „Voksie". Odkrycie, że to on jest moim sobowtórem, właściwie w niczym nie zmieniło mojego życia. Mimo to jest naprawdę dość fajnie poznać najbardziej podobną do ciebie osobę na świecie, zwłaszcza jeżeli jest to ktoś, kogo podziwiasz. Gdy skończę pisać tę książkę i przestanę prowadzić pustelnicze życie, to może się spotkam z Matthewsem osobiście i pogadamy sobie o publicystyce Jamesa Surowieckiego.

Poszukiwania sobowtóra Ortiza były gratką dla kibiców bejsbola, a poszukiwania mojego sobowtóra przyjemne, w każdym razie dla mnie. Czy mogą one jednak przynieść coś więcej? Z tego rodzaju metody korzystały na przykład duże firmy internetowe, by znacznie uatrakcyjnić swoją ofertę i lepiej poznać klientów. Amazon używa czegoś w rodzaju „metody na sobowtóra", żeby zasugerować, jakie książki moglibyście chcieć u nich kupić. Patrzą, co wybierają ludzie do was podobni, i na tym opierają swoje rekomendacje.

To samo robi Pandora, sugerując, jakich może chcielibyście posłuchać piosenek. Na podobnej zasadzie Netflix domyśla się, jakie filmy mogą wam się podobać. Wyniki zastosowania tych technik były tak udane, że kiedy główny informatyk Amazona, Greg Linden, wprowadził „metodę na sobowtóra", by prognozować preferencje czytelnicze, zamówienia firmy wzrosły na tyle, że jej założyciel Jeff Bezos padł na kolana i zawołał do Lindena: „Panie, nie jestem godzien!".

Naprawdę ciekawe w „metodzie na sobowtóra", zważywszy na jej moc, jest jednak nie to, jak powszechnie jest dziś używana, tylko jak często używana nie jest. A przecież dzięki tego rodzaju personalizacji, na jaką pozwala, można by znacznie poprawić różne sfery życia. Weźmy na przykład zdrowie.

Isaac Kohane, informatyk i analityk medyczny z Harvardu, próbuje zastosować „metodę na sobowtóra" w medycynie. Chce zebrać i uporządkować wszystkie informacje o pacjencie, by zamiast podchodzić do niego według ogólnego wzorca, przykładanego do wszystkich chorych, lekarze mogli znaleźć takich pacjentów jak on – a zatem, żeby mogli stawiać bardziej osobiste i dokładniejsze diagnozy i stosować bardziej spersonalizowane kuracje.

Kohane uważa, że jest to naturalne rozszerzenie pola działania dla nauk medycznych, a przy tym wcale nie nazbyt radykalne. „Co to jest diagnoza? – pyta. – Właściwie jest to stwierdzenie, które oznacza, że dzielisz pewne cechy z innymi wcześniej badanymi pacjentami. Gdybym, co nie daj Bóg, zdiagnozował u ciebie atak serca, powiedziałbym, że wykryłem stan patofizjologiczny, który, o czym wiem na podstawie badań innych pacjentów, oznacza, że przeszedłeś zawał mięśnia sercowego"[20].

Diagnoza medyczna jest więc w istocie wynikiem zastosowania prymitywnej „metody na sobowtóra". Problem polega na tym, że zbiory danych, których używają lekarze, stawiając rozpoznanie, są niewielkie. Lekarz opiera się na własnym doświadczeniu zawodowym, czyli na znajomości pacjentów, których leczył osobiście, i być może na wiedzy akademickiej, jaką czerpie z artykułów innych lekarzy lub badaczy, również mających do czynienia z małymi grupami pacjentów. Jak jednak widzieliśmy, żeby „metoda na pacjenta" naprawdę się sprawdziła, potrzebujemy znacznie więcej przypadków.

Big data na tym polu faktycznie może pomóc. Dlaczego zatem wprowadzenie tej metody do medycyny tak długo trwa? Dlaczego nie stosuje się jej powszechnie? Problem polega na zebraniu danych. Większość historii chorób czy kart pacjentów wciąż pozostaje na papierze w tekturowych teczkach, a z kolei wielu „skomputeryzowanych" lekarzy przechowuje takie dane w niekompatybilnych formatach. Kohane twierdzi, że często mamy lepsze dane o bejsbolu niż o zdrowiu. Ale nawet proste środki mogą nas daleko zaprowadzić – tu Kohane często podaje obrazowy przykład „nisko wiszących", czyli łatwo dostępnych „owoców". Uważa między innymi, że już samo stworzenie kompletnego zbioru danych,

zawierającego wzrost, wagę i ewentualne schorzenia wszystkich amerykańskich dzieci, miałoby rewolucyjne znaczenie dla pediatrii, bo tempo wzrostu jednego dziecka można by porównać z tempem wzrostu jego rówieśników. Komputer wyszukałby dzieci rozwijające się z podobną szybkością i automatycznie oznaczyłby wszystkie potencjalnie niepokojące zjawiska. Na przykład wykrywałby, że dziecko za wcześnie przestało rosnąć, co w pewnych sytuacjach może oznaczać dwie przypadłości: niedoczynność tarczycy albo guz mózgu, a wczesna diagnoza w obu sytuacjach byłaby darem od losu. „To rzadkie przypadki, jeden na dziesięć tysięcy – twierdzi Kohane. – Dzieci w większości są zdrowe. Ale dzięki temu można by je diagnozować wcześniej, przynajmniej o rok. Dałoby się to zrobić, na sto procent".

James Heywood, przedsiębiorca, nieco inaczej podchodzi do trudności z gromadzeniem danych medycznych[21]. Stworzył portal PatientsLikeMe.com, gdzie ludzie mogą umieszczać informacje o sobie – na co chorują, jak się leczą i jakie odczuwają skutki uboczne. Udało mu się dokonać już bardzo wiele, na przykład stworzyć wykresy różnych przebiegów chorób i przedstawić je na tle naszej potocznej wiedzy na ich temat.

Jego celem jest zaangażowanie w ten projekt tak wielu osób i opisanie tylu chorób, by ludzie mogli znaleźć swojego, już wyleczonego, sobowtóra. Heywood ma nadzieję, że dzięki niemu znajdziecie ludzi w swoim wieku i swojej płci, z podobną historią choroby, którzy mieli podobne objawy, i dowiecie się, co im pomogło. Gdyby to się udało, medycyna naprawdę zupełnie zmieniłaby swoje oblicze.

OPOWIEŚCI Z BAZY DANYCH

Jeżeli chodzi o mnie, „filmowe" zbliżenie na mały wycinek danych uważam za bardziej wartościowe niż konkretne wyniki pojedynczego badania, bo dzięki uszczegółowieniu możemy w nowy sposób spojrzeć na nasze życie i mówić o nim inaczej niż dotychczas.

Kiedy ludzie dowiadują się, że jestem analitykiem i piszę książki o swojej pracy, czasem dzielą się ze mną znajomością jakichś faktów czy sondaży. Ale te dane na ogół wydają mi się nudne – statyczne i bez życia. Brak w nich ducha opowieści.

Znajomi próbowali mnie też wciągnąć w lekturę beletrystyki i biografii, one też mnie jednak nie ciekawią. Zawsze stawiam sobie przy nich pytanie: „Czy to mogłoby się zdarzyć w innych sytuacjach? Czy istnieje zasada natury bardziej ogólnej?". Historie literackie wydają mi się mało ważne i niereprezentatywne.

Starałem się pokazać w tej książce coś, co według mnie nie da się z niczym porównać, a co opiera się na danych i liczbach, jest pouczające i dalekosiężne. Mimo to dzisiejsze zbiory danych są tak bogate, że możemy sobie wyobrazić kryjących się pod nimi prawdziwych ludzi. Kiedy przyglądamy się zużyciu wody w Edmonton minuta po minucie, to widzę w wyobraźni, jak pod koniec tercji kibice wstają z kanap. Kiedy słyszymy, że ludzie po przeprowadzce z Filadelfii do Miami zaczynają oszukiwać państwo na podatkach, to widzę ich, jak rozmawiają z sąsiadami i dowiadują się, na czym polega opisany przeze mnie trik podatkowy. A gdy przywołujemy fanów bejsbola w dowolnym wieku, widzę swoje dzieciństwo, dzieciństwo mojego brata i miliony już dorosłych ludzi, płaczących nad drużyną, która zdobyła ich serca, kiedy mieli mniej więcej osiem lat.

Raz jeszcze ryzykując, że zabrzmi to patetycznie, powiem, że ekonomiści i analitycy wspomniani w mojej pracy tworzą nie tylko nowe narzędzie poznania, ale też nowy gatunek literacki. W tym rozdziale, tak jak i w znacznej części książki *Wszyscy kłamią*, starałem się pokazać zbiór danych tak obszerny, bogaty i pozwalający nam tak dokładnie przyjrzeć się jego wybranemu fragmentowi, że chociaż nie ograniczamy się do żadnej konkretnej, niereprezentatywnej jednostki ludzkiej, możemy opowiadać barwne i złożone historie o człowieku.

ROZDZIAŁ SZÓSTY
ŚWIAT TO JEDNO WIELKIE LABORATORIUM

Dwudziesty siódmy dzień lutego 2000 roku rozpoczął się na google'owskim kampusie Mountain View jak każdy inny[1]. Świeciło słońce, rowerzyści pedałowali, masażyści masowali, a pracownicy firmy nawadniali się sokiem ogórkowym. Ale to właśnie wtedy, tego zwyczajnego dnia, kilku pracowników Google'a wpadło na pomysł obejmujący sekret, który napędza dzisiaj większość Internetu. Informatycy znaleźli najlepszy sposób na to, żebyście klikali myszką, wracali na pewne portale i długo na nich siedzieli.

Zanim opiszę, co zrobili, musimy porozmawiać o różnicy między korelacją i przyczynowością, jest to bowiem niezwykle ważna kwestia w analizie danych – i to taka, która nie została jeszcze odpowiednio zbadana.

Media codziennie bombardują nas badaniami opartymi na korelacjach. Mówi się nam chociażby, że ludzie spożywający alkohol w niewielkich ilościach są na ogół zdrowsi. To przykład korelacji.

Czy to znaczy, że niewielkie ilości alkoholu mogą korzystnie wpłynąć na nasze zdrowie – czyli że w grę wchodzi tutaj przyczynowość? Być może nie. Możliwe, że to dobre zdrowie sprawia, że ludzie czasami piją. Taką przyczynowość socjologowie nazywają odwróconą. Możliwe też, że istnieje jakiś niezależny czynnik, który z jednej strony skłania nas do spożywania alkoholu, a z drugiej wpływa pozytywnie na stan naszego organizmu. Niewykluczone, że częste spędzanie czasu z przyjaciółmi prowadzi zarówno do konsumpcji napojów wyskokowych, jak

i poprawia nam zdrowie. Socjologowie nazywają to sytuacją opuszczonej zmiennej.

Jak zatem możemy lepiej wykrywać przyczynowość? Najskuteczniejszą metodą jest stochastyczny, ale kontrolowany eksperyment, który przeprowadza się w następujący sposób: najpierw dzielimy ludzi losowo na dwie grupy. Jedna, tak zwana grupa robocza, proszona jest o to, żeby coś robiła albo zażywała jakiś preparat, druga, tak zwana grupa kontrolna, nie. Następnie obserwujemy, co dzieje się z członkami obu grup, a różnica w wynikach obserwacji to efekt przyczynowy.

Na przykład żeby zbadać, czy spożywanie niewielkich ilości alkoholu faktycznie przyczynia się do poprawy stanu zdrowia, możecie losowo wybrać kilka osób i poprosić, żeby przez rok piły po kieliszku wina dziennie. Potem zwróćcie się do kilku innych osób, żeby przez rok w ogóle nie spożywały alkoholu, a następnie porównajcie stan zdrowia obu grup. Ponieważ wasi badani zostali wybrani losowo, nie ma powodu przypuszczać, że pierwsza grupa będzie początkowo w lepszym stanie zdrowia niż druga albo że kiedyś bardziej udzielała się towarzysko. Z tego względu można sądzić, że skutki spożycia wina są kauzalne. Takie kontrolowane losowe eksperymenty dostarczają najbardziej godnych zaufania wyników w każdej dziedzinie. Jeżeli na przykład jakieś lekarstwo pomyślnie przejdzie tego rodzaju próbę, można je podawać pacjentom. Jeżeli nie, nie trafi na półki apteczne.

Eksperymentów losowych używa się też coraz częściej w naukach społecznych. Esther Duflo, francuska ekonomistka z MIT, prowadzi kampanię na rzecz częstszego stosowania ich w ekonomii rozwoju, dziedzinie nauki, która próbuje ustalić, w jaki sposób najlepiej jest pomagać najuboższym. Zastanówcie się nad badaniami Duflo i jej kolegów, usiłujących poprawić wykształcenie na wiejskich obszarach Indii, gdzie ponad połowa uczniów szkół średnich nie potrafi przeczytać prostego zdania. Jednym z możliwych powodów takiego stanu rzeczy jest to, że nauczyciele nie przychodzą regularnie do pracy. W niektórych indyjskich szkołach każdego dnia przeszło czterdzieści procent nauczycieli jest nieobecnych.

Na czym polegały badania Duflo? Podzieliła losowo szkoły na dwie grupy. Członkowie pierwszej z nich (grupy roboczej) oprócz pensji podstawowej dostawali niewielką kwotę dodatkową – pięćdziesiąt rupii, czyli około dolara i piętnastu centów – za każdy dzień, który spędzali w pracy. W drugiej grupie, kontrolnej, nie płacono nadprogramowo za obecność. Rezultaty były zdumiewające. W grupie wynagradzanej dodatkowo absencja nauczycieli spadła o połowę[2]. Znacznie poprawiły się też wyniki uczniów na sprawdzianach i testach, zwłaszcza wśród dziewcząt. Pod koniec eksperymentu dziewczynki w tych szkołach, w których pedagogom płacono ponad plan, miały o siedem punktów procentowych więcej szans, żeby nauczyć się pisać.

Jak doniósł „New Yorker", eksperymenty Esther Duflo zrobiły ponoć takie wrażenie na Billu Gatesie[3], że powiedział jej: „Po prostu musimy sponsorować twoje badania, i już".

ABECADŁO TESTÓW A/B

Widzimy więc, że eksperymenty losowe to najlepsza metoda wykrywania przyczynowości, a ich zastosowanie rozpowszechniło się w naukach społecznych – co prowadzi nas z powrotem do biur Google'a i dnia 27 lutego 2000 roku. Co takiego zrobił wtedy Google, że zrewolucjonizował Internet?

Otóż kilku informatyków postanowiło przeprowadzić podobny eksperyment na portalu internetowym i podzieliło użytkowników na dwie grupy. Grupie roboczej pokazywano dwadzieścia linków na stronie z wynikami wyszukiwań, a grupie kontrolnej tylko dziesięć, jak zwykle. Następnie informatycy porównali poziom zadowolenia obu grup, mierząc go liczbą powrotów do wyszukiwarki.

To ma być rewolucja? Ten eksperyment nie wygląda zbyt rewolucyjnie. Wspominałem już, że takie losowe doświadczenia prowadzą również firmy farmaceutyczne i socjologowie, więc o co tyle krzyku?

Najbardziej istotna kwestia, i to informatycy Google'a szybko sobie uświadomili, polega na tym, że eksperymenty w świecie cyfrowym mają ogromną przewagę nad eksperymentami offline. Te ostatnie mogą być przekonujące, ale wymagają sporych nakładów. Do badań prowadzonych przez Duflo trzeba było skontaktować się ze szkołami, zdobyć fundusze, zapłacić przynajmniej niektórym nauczycielom i sprawdzić postępy wszystkich uczniów. Eksperymenty offline mogą kosztować tysiące albo setki tysięcy dolarów i ciągnąć się miesiącami albo nawet latami.

W świecie cyfrowym eksperymenty stochastyczne mogą być tanie i szybkie. Nie trzeba specjalnie rekrutować uczestników ani im płacić. Wystarczy napisać składający się z jednej linijki program, który losowo podzieli ich na grupy. Nie będą musieli wypełniać żadnych ankiet, zamiast tego wystarczy policzyć liczbę kliknięć i posunięć myszką. Nie trzeba samodzielnie odkodowywać ani analizować odpowiedzi, można napisać program, który automatycznie zrobi to za ciebie. Nie musisz się też z nikim kontaktować. Nie musisz nawet mówić użytkownikom, że biorą udział w jakimś eksperymencie.

To czwarta moc zbioru big data: można z jego pomocą przeprowadzać losowe eksperymenty i dzięki nim odkrywać prawdziwą przyczynowość. Eksperymenty te są o wiele, wiele łatwiejsze w realizacji, można je prowadzić w każdym czasie i właściwie wszędzie, byle online. W epoce big data świat stał się jednym wielkim laboratorium.

Nowinka ta szybko rozeszła się po firmie, a potem po całej Dolinie Krzemowej, w której kontrolowane losowe eksperymenty nazwano „testami A/B". W 2011 roku informatycy Google'a przeprowadzili ich aż siedem tysięcy[4], a robi się tych testów coraz więcej.

Gdy Google chce się dowiedzieć, w jaki sposób skłonić więcej osób do klikania w reklamy, może spróbować na przykład wykorzystać w nich dwa odcienie błękitu – jeden dla grupy A, drugi dla B – a następnie porównać liczbę kliknięć. Oczywiście łatwość przeprowadzania takich testów skutkuje czasem ich zbyt częstym wykorzystywaniem. Niektórzy pracownicy firmy uważali, że ponieważ eksperymenty A/B nie wymagają właściwie żadnego wysiłku, Google korzysta z nich

z nadmierną częstotliwością. W 2009 roku pewien sfrustrowany grafik zrezygnował z pracy, gdy firma wypróbowała w tego typu testach czterdzieści jeden nieznacznie różniących się od siebie odcieni błękitu[5]. Sprzeciw grafika, który przedkładał sztukę ponad obsesyjne badania rynku, na dobrą sprawę ani trochę nie powstrzymał jednak rozpowszechniania opisywanej metody.

Facebook przeprowadza obecnie tysiąc testów typu A/B dziennie[6], co oznacza, że niewielka liczba informatyków każdego dnia przygotowuje więcej kontrolowanych losowych eksperymentów niż cały amerykański przemysł farmaceutyczny w ciągu roku.

Ale z takich eksperymentów korzystają już nie tylko największe firmy komputerowe. Były pracownik Google'a, Dan Siroker, zastosował tę nową metodę badawczą w czasie pierwszej kampanii prezydenckiej Baracka Obamy, wykorzystując ją do badania projektów stron internetowych, e-maili z propozycjami handlowymi [*e-mail pitches*] i formularzy dotacji. Po jakimś czasie założył własną firmę, Optimizely[7], która umożliwia różnym organizacjom i instytucjom przeprowadzanie szybkich testów A/B. W 2012 roku z usług Optimizely korzystał Barack Obama, podobnie jak jego przeciwnik, Mitt Romney, chcąc zdobyć więcej podpisów, wolontariuszy i dotacji. Sirokera zatrudniają też firmy tak różne, jak Netflix i TaskRabbit czy czasopismo „New York".

Żeby się przekonać, jak bardzo wartościowy może być test A/B, zobaczmy, w jaki sposób wykorzystał go Obama, aby zaangażować w swoją kampanię wyborczą możliwie najwięcej osób. Strona internetowa przyszłego prezydenta składała się początkowo z jego zdjęcia wraz z hasłem: „Zaangażuj się" i usytuowanego pod nim okienka z przyciskiem: „Zarejestruj się", zachęcającego internautów, żeby zarejestrowali się na stronie, popierając w ten sposób kandydata demokratów.

Czy był to najlepszy sposób powitania ewentualnych zwolenników? Z pomocą Optimizely zespół Baracka Obamy zaczął sprawdzać, czy inne zdjęcie i okienko nie skłonią do rejestracji większej liczby ludzi. Czy będą klikali chętniej, jeżeli na stronie pojawi się fotografia Obamy z bardziej poważną miną? A może więcej ludzi się zarejestruje, jeśli tekst w okienku będzie głosił: „Dołącz do nas już dziś”? Zespół doradców pokazywał użytkownikom różne kombinacje zdjęć i okienek i sprawdzał, ilu z nich w nie kliknęło, żeby się zarejestrować. Ciekawe, czy zgadniecie, które zdjęcie i jaki napis okazały się najskuteczniejsze…

Testowane zdjęcia

Testowane przyciski

Zwyciężyły fotografia rodziny Obamy opatrzona hasłem: „Zmiana, w którą możemy uwierzyć” i przycisk z napisem: „Dowiedz się więcej”. I to w cuglach. Jak wyliczyli doradcy, dzięki tej kombinacji kandydata demokratów poparło o czterdzieści procent więcej osób, a na kampanię wyborczą zebrano dodatkowe sześćdziesiąt milionów dolarów[8].

Zwycięska kombinacja

Jest jeszcze jedna wielka korzyść z tego, że tę najlepszą z metod testowania można stosować takim tanim kosztem i z taką łatwością: otóż dzięki niej nie musimy polegać na intuicji, która, jak powiedzieliśmy w rozdziale pierwszym, ma swoje ograniczenia. Testy typu A/B okazują się tak ważne, ponieważ ludzie są nieprzewidywalni i intuicyjnie często nie jesteśmy w stanie przewidzieć, jak się zachowają.

A czy intuicja podpowiedziała wam, które połączenie zdjęcia i napisu okazało się optymalne dla Obamy?

Możecie wystawić ją teraz na próbę ponownie. Oto nagłówki z gazety „Boston Globe”[9] – spróbujcie odgadnąć, który człon z dziewięciu podanych par uzyskał więcej kliknięć:

Jeden z poniższych nagłówków uzyskał o wiele więcej kliknięć niż drugi

	Nagłówek A	**Nagłówek B**
1.	Czy dron SnotBot uratuje wieloryby?	Czy ten dron może pomóc uratować wieloryby?
2.	Oczywiście „zgniłe jaja" to najczęściej wyszukiwane hasło w Massachusetts	Najczęściej wyszukiwane hasło w Mass. jest mocno żenujące
3.	Zawody w ciupcianiu i proces o gwałt w St. Paul	Nie postawiono zarzutów w seksualnym skandalu w podstawówce
4.	Kobieta rozbija bank dzięki rzadkiej karcie bejsbolowej	Kobieta dostaje 179 000 dolarów za rzadką kartę bejsbolową
5.	Roczny deficyt MBTA podwoi się do 2020 roku	Przygotujcie się: deficyt MBTA wkrótce się podwoi
6.	Jak Massachusetts pomógł ci uzyskać prawo dostępu do środków antykoncepcyjnych	Jak Boston University pomógł zlikwidować „przestępstwa przeciwko czystości płciowej"
7.	Kiedy w Bostonie otwarto pierwszą linię metra	Rysunkowa historia otwarcia pierwszej linii metra w Bostonie
8.	Ofiara i rodzina oskarżają toksyczną kulturę w procesie o gwałt w podstawówce	Ofiara i rodzina wydają oświadczenie w procesie o gwałt w podstawówce
9.	Facet w czapeczce z napisem „Free Brady" jako jedyny psuje figiel Miley Cyrus	Fan Patsów zaskoczony widokiem Miley Cyrus w przebraniu

Skończyliście? Poprawne odpowiedzi zostały wytłuszczone poniżej.

	Nagłówek A	**Nagłówek B**	**Zwycięzca**

1.	**Czy dron SnotBot uratuje wieloryby?**	Czy ten dron może pomóc uratować wieloryby?	53% więcej kliknięć dla A
2.	Oczywiście „zgniłe jaja" to najczęściej wyszukiwane hasło w Massachusetts	**Najczęściej wyszukiwane hasło w Mass. jest mocno żenujące**	986% więcej kliknięć dla B
3.	Zawody w ciupcianiu i proces o gwałt w St. Paul	**Nie postawiono zarzutów w seksualnym skandalu w podstawówce**	108% więcej kliknięć dla B
4.	**Kobieta rozbija bank dzięki rzadkiej karcie bejsbolowej**	Kobieta dostaje 179 000 dolarów za rzadką kartę bejsbolową	38% więcej kliknięć dla A
5.	Roczny deficyt MBTA podwoi się do 2020 roku	**Przygotujcie się: deficyt MBTA wkrótce się podwoi**	62% więcej kliknięć dla B
6.	Jak Massachusetts pomógł ci uzyskać prawo dostępu do środków antykoncepcyjnych	**Jak Boston University pomógł zlikwidować „przestępstwa przeciwko czystości płciowej"**	188% więcej kliknięć dla B
7.	**Kiedy w Bostonie otwarto pierwszą linię metra**	Rysunkowa historia otwarcia pierwszej linii metra w Bostonie	33% więcej kliknięć dla A
8.	Ofiara i rodzina oskarżają toksyczną kulturę w procesie o gwałt w podstawówce	**Ofiara i rodzina wydają oświadczenie w procesie o gwałt w podstawówce**	76% więcej kliknięć dla B
9.	Facet w czapeczce z napisem „Free Brady" jako jedyny psuje figiel Miley Cyrus	**Fan Patsów zaskoczony widokiem Miley Cyrus w przebraniu**	67% więcej kliknięć dla B

Myślę, że odgadliście poprawnie ponad połowę, być może odpowiadając po prostu tak, jak sami byście klikali. Ale zapewne nie domyśliliście się wszystkich odpowiedzi.

Dlaczego? Co wam umknęło? Czego nie wiecie o ludzkim behawiorze? Czego możecie się nauczyć na podstawie swoich pomyłek?

Takie pytania zadajemy sobie zwykle po nietrafionych prognozach.

Popatrzcie jednak, jak trudno jest wyciągnąć jakieś ogólne wnioski z nagłówków „Globe'a". W pierwszym teście wystarczyło zmienić właściwie jedno słowo, „ten", na „SnotBot", aby uzyskać znacznie więcej kliknięć. Mogłoby z tego wynikać, że zwycięża większa liczba szczegółów – ale detaliczne „zgniłe jaja" w drugim nagłówku przegrywają. W nagłówku czwartym „rozbija bank" wygrywa z sumą 179 000 dolarów. To z kolei mogłoby sugerować, że zwycięża idiomatyka albo slang. Ale określenie „zawody w ciupcianiu" w trzecim nagłówku przegrywa.

Nauka płynąca z testów typu A/B w znacznym stopniu brzmi tak, żeby ostrożnie podchodzić do uogólnień. Clark Benson jest dyrektorem ranker.com, portalu informacyjnego i rozrywkowego, który przy wyborze nagłówków i projektów stron w znacznym stopniu polega na testach A/B. „W ostatecznym rozrachunku nie można niczego zakładać z góry – twierdzi. – Trzeba sprawdzać dosłownie wszystko"[10].

Eksperymenty wypełniają takie luki w naszej wiedzy o naturze człowieka, które istnieją i będą istnieć zawsze, bo ciągle powstają nowe. Gdybyśmy wiedzieli, jak je uzupełnić, opierając się wyłącznie na naszym doświadczeniu życiowym, testy online nie miałyby żadnej wartości. Ale nie wiemy, więc ją mają.

Kolejny powód, dla którego eksperymenty typu A/B są bardzo ważne, jest taki, że wykrywają z pozoru niewielkie zmiany, które mogą przynieść poważne skutki. Jak to ujmuje Benson: „Nie przestaje mnie zdumiewać, że drobne, naprawdę bardzo drobne czynniki mają tak ogromną wartość testową".

W grudniu 2012 roku Google zmienił formę publikowanych u siebie reklam, dodając do nich kwadracik ze skierowaną w prawo strzałką[11].

Spójrzcie, jaka kuriozalna jest ta strzałka. Nie pokazuje absolutnie nic. Gdy strzałki pojawiły się po raz pierwszy, krytykowało je wielu użytkowników Google'a[12]. Zastanawiali się, po co firma dodaje do reklam nic nieznaczący symbol.

Cóż, Google chroni swoje tajemnice biznesowe, nie mówi więc, jaką dokładnie wartość miały te oznaczenia – pracownicy firmy przyznali jednak, że zaczęto je wykorzystywać, bo zwyciężyły w teście typu A/B, a dzięki nim o wiele więcej osób klikało w reklamy. Ta drobna, z pozoru nic niewnosząca zmiana przyniosła Google'owi i jego partnerom reklamowym mnóstwo forsy.

Jak się znajduje tego rodzaju drobiazgi, które przynoszą potem wielkie zyski? Najpierw trzeba przetestować bardzo dużo wariantów, nawet, wydawałoby się, bardzo trywialnych. Użytkownicy Google'a widzieli wielokrotnie, że forma reklam w wyszukiwarce zmienia się nieznacznie i szybko wraca do poprzedniego kształtu. Mimo woli stali się więc członkami grup roboczych w testach A/B – ale ich jedynym wynagrodzeniem były właśnie tylko te drobne zmiany.

Eksperyment z wycentrowaniem reklamy (nieudany)

Najlepiej sprzedający się Ipad 2 Case
The ZAGGmate™ — mocna aluminiowa obudowa
z klawiaturą Bluetooth
www.zagg.com

Eksperyment z zielonymi gwiazdkami (nieudany)

Recenzje restauracji Foster's Hollywood, Madryt, Hiszpania…
www.tripadvisor.co.uk >… > Madryt > restauracje w Madrycie ▾ TripAdvisor ▾
★★★☆☆ Ocena 3 — 118 recenzji
Foster's Hollywood, Madryt. Zob. 118 **bezstronnych** recenzji restauracji **Foster's Hollywood**, rating 3 gwiazdki na 5 pięć możliwych na TripAdvisor i 3647 miejsce na 6489 restauracji w Madrycie.

Eksperyment z nowym krojem pisma (nieudany)

Wiadomości giełdowe na żywo
Darmowe wykresy, informacje i wskazówki od ekspertów UTVi.
Odwiedźcie nas już dziś!
UTVi.com/Stocks

Powyższe warianty reklam nie trafiły na szeroki rynek. Przegrały. Stanowiły jednak część procesu, dzięki któremu można było wybrać zwycięzców. Droga do strzałki o dużej klikalności wybrukowana jest brzydkimi gwiazdkami, nieszczęśliwym usytuowaniem tekstu i jajcarską czcionką.

Fajnie jest zgadywać, co sprawia, że ludzie klikają w jakieś okienko. A jeżeli należysz do zwolenników demokratów, miło ci będzie wiedzieć, że dzięki testom A/B Obama zebrał więcej pieniędzy. Testy te mają też jednak ciemną stronę.

W swojej znakomitej książce *Uzależnienia 2.0* Adam Alter[13] pisze o tym, jak we współczesnym społeczeństwie powstają uzależnienia behawioralne. Wielu z nas jest coraz trudniej żyć bez korzystania z pewnych funkcji Internetu.

Mój ulubiony zbiór danych, wyszukiwarka Google'a, może nam podpowiedzieć, od czego jesteśmy najbardziej uzależnieni. Otóż wydaje się, że w większości wypadków chodzi o te same nałogi, z którymi borykamy się od dawna – na przykład narkotyki, seks i alkohol. Lecz na tej liście pojawia się już Internet, chociażby w postaci pornografii i Facebooka.

Osiem najpowszechniejszych uzależnień według wpisów w Google'u, 2016[14]

Narkotyki	Alkohol	Hazard
Seks	Cukier	Facebook
Pornografia	Miłość	

I należy przyznać, że to między innymi testy typu A/B mogą sprawiać, że Internet stał się tak cholernie uzależniający.

W *Uzależnieniach 2.0* Alter powołuje się na Tristana Harrisa, etyka designu, który tłumaczy, dlaczego nie potrafimy się oprzeć niektórym portalom w sieci: bo po drugiej stronie ekranu siedzi tysiąc osób, których zadaniem jest właśnie złamać nasz opór.

A ludzie ci stosują testy A/B.

To dzięki nim Facebook wie, że nadanie jakiemuś okienku konkretnego koloru skłania użytkowników, by częściej wracali na stronę – dlatego zmieniają kolory. Pracownicy portalu mogą też zauważyć, że do powrotu na stronę zachęca nas na przykład określony krój pisma – dlatego go zmieniają. I może się okazać, że wysyłanie e-maili o określonej porze dnia skłania internautów do częstszego odwiedzania strony – dlatego piszą do nich właśnie w tych godzinach.

Facebook wkrótce stanie się portalem badającym przede wszystkim to, ile czasu na nim spędzamy. Innymi słowy, tam, gdzie znajdziemy wystarczająco dużo zwycięzców testów A/B, mamy portal uzależniający. Tego rodzaju feedbacku ze strony klientów nie dostawały nigdy nawet firmy tytoniowe.

Testy A/B coraz częściej stają się też narzędziem używanym przez producentów gier komputerowych. Jak pisze Alter, *World of Warcraft* testuje w ten sposób różne wersje swoich rozgrywek. W jednej twoim zadaniem może być zabicie jakiejś postaci, w drugiej uratowanie kogoś albo czegoś. Projektanci gier mogą podsuwać ludziom na próbę różne propozycje i sprawdzać, które zdobywają większą popularność. Może się na przykład okazać, że do rozgrywki, w której twoim zadaniem było kogoś ocalić, użytkownicy wracali o trzydzieści procent częściej. Im więcej wariantów sprawdzają projektanci, tym więcej znajdują między nimi zwycięzców. Trzydzieści procent tu, trzydzieści procent tam, aż w końcu mają grę, nad którą wielu dorosłych mężczyzn ślęczy nawet w tej chwili w piwnicy domu rodziców.

Jeżeli trochę was to niepokoi, to przyznam, że mnie też. Pod koniec książki pomówimy jeszcze nieco na temat etycznych implikacji tego zjawiska i innych aspektów big data. Ale bez względu na skutki, takie eksperymenty to dzisiaj podstawowe narzędzie w warsztacie analityka danych. Korzysta on jednak także z jeszcze innej formy eksperymentu, próbując odpowiedzieć na bardzo różne pytania, jak na przykład, czy reklama telewizyjna jest naprawdę skuteczna.

OKRUTNE – CHOĆ POUCZAJĄCE – EKSPERYMENTY NATURY

Jest 22 stycznia 2012 roku, New England Patriots grają przeciwko Baltimore Ravens w meczu o mistrzostwo AFC.

Do końca pozostała jedna minuta. Drużyna Baltimore przegrywa, ale ma piłkę. Następne sześćdziesiąt sekund zdecyduje, który zespół zagra w Super Bowl. Następne sześćdziesiąt sekund zapewni graczom sławę.

Ostatnia minuta meczu przyczyni się też jednak do czegoś, co przynajmniej dla ekonomisty jest o wiele ważniejsze: dowiemy się bowiem raz na zawsze, czy reklamy telewizyjne są rzeczywiście skuteczne.

Założenie, że reklama zwiększa sprzedaż, jest oczywiście fundamentalne dla każdej gospodarki, ale diablo trudne do udowodnienia. Mamy tu do czynienia z podręcznikowym przykładem trudności w rozróżnieniu korelacji i przyczynowości.

Wiemy na pewno, że najlepiej sprzedają się produkty reklamowane najczęściej. Wytwórnia Twentieth Century Fox wydała sto pięćdziesiąt milionów dolarów na wypromowanie filmu *Avatar*, który stał się najlepiej zarabiającym filmem wszech czasów. Ale ile z dwóch i pół miliarda dolarów, które zarobił ten obraz, zawdzięczamy agresywnej reklamie? Wytwórnia przynajmniej po części dlatego wydała tyle pieniędzy na promocję, że wiedziała, iż ma w ręku interesujący produkt.

Przedsiębiorcy uważają, że wiedzą, jak skuteczne są ich reklamy, lecz ekonomiści pozostają co do tego sceptyczni. Profesor Steven Levitt z University of Chicago współpracował kiedyś z pewną firmą elektroniczną i nie mógł się nadziwić, gdy jej pracownicy próbowali go przekonać, że wiedzą, jak bardzo skuteczne są reklamy ich produktów. Zastanawiał się, skąd czerpią taką pewność.

Wyjaśniono mu, że co roku przed Dniem Ojca zwiększają nakłady na reklamę w telewizji. I oczywiście co roku przed Dniem Ojca sprzedają najwięcej produktów. Hmm, ale to może dlatego, że mnóstwo dzieci kupuje elektronikę dla swoich ojców, zwłaszcza w prezencie na ich święto, i to bez względu na reklamę.

„Ci ludzie kompletnie opacznie rozumieją przyczynowość", oświadczył Levitt na jednym ze swoich wykładów[15]. Oczywiście możliwe, że tak jest. Nie wiemy tego na pewno. „To naprawdę twardy orzech do zgryzienia", dodał.

Problem wydaje się niezwykle ważny, ale przedsiębiorstwa niechętnie przeprowadzają eksperymenty, które mogą pomóc go rozwiązać. Levitt próbował przekonać swoją firmę, żeby zgodziła się na kontrolowany losowy eksperyment i sprawdziła dokładnie, jak

skuteczna jest jej reklama w TV. Ponieważ testy typu A/B nie są jeszcze możliwe w telewizji, trzeba byłoby zobaczyć, co się dzieje ze sprzedażą niektórych produktów, jeśli w ogóle pozbawić je reklamy.

Ale przedstawiciele spółki zareagowali tak: „Czy pan oszalał? Nie możemy się nie reklamować na dwudziestu rynkach. Dyrektor nas zabije".

Na tym Levitt zakończył współpracę z firmą.

I tu wracamy do meczu Patriots kontra Ravens. W jaki sposób wynik meczu futbolowego może pomóc określić przyczynowe skutki reklamy? Na pewno nie powie nam, jakie efekty przyniosła konkretna kampania reklamowa konkretnej firmy, może jednak dostarczyć informacji na temat średniej skuteczności reklam z wielu różnych kampanii.

Wychodzi na to, że przy okazji tego rodzaju zawodów sportowych mamy do czynienia z czymś w rodzaju ukrytego eksperymentu reklamowego. Oto na czym on polega: firmy kupują i produkują reklamy na Super Bowl przed rozpoczęciem rozgrywek, ale gdy decydują, które reklamy nadawać, nie wiedzą jeszcze przecież, jakie drużyny będą grały w finale.

A wyniki playoffów mają ogromny wpływ na to, kto obejrzy Super Bowl. Dwie drużyny, które ostatecznie wejdą do finału, przyciągną przed ekrany olbrzymią publiczność. Jeżeli wygra New England, zespół, który rozgrywa swoje mecze pod Bostonem, to finał zobaczy znacznie więcej osób w Bostonie niż w Baltimore. I vice versa.

Dla reklamodawców odpowiedź na pytanie o to, czy dodatkowe dziesiątki tysięcy widzów obejrzą ich reklamy raczej w Baltimore, czy raczej w Bostonie, przypomina rzut monetą, i to wykonany już po tym, jak ich reklamy zostały zakupione i wyprodukowane.

Wróćmy na boisko, gdzie Jim Nantz ze stacji telewizyjnej CBS ogłasza końcowy wynik meczu:

> Biegnie Billy Cundiff, będziemy mieli remis i najprawdopodobniej dogrywkę. W meczach rozgrywanych przez ostatnie dwa lata byłoby więc w sumie szesnaście do szesnastu. Trzydzieści dwa jardy i mamy remis. Jest kopnięcie. Uwaga! Uwaga!

> Ale niecelne… A zatem to Patriots są górą i to oni pojadą do Indianapolis. To oni zagrają w czterdziestym szóstym finale Super Bowl.

Dwa tygodnie później w XLVI finale Super Bowl publiczność w Bostonie miała sześćdziesiąt i trzy dziesiąte procent udziałów oglądalności, a w Baltimore pięćdziesiąt i dwie dziesiąte. Reklamy w Bostonie obejrzało sześćdziesiąt tysięcy osób więcej.

Rok później te same dwie drużyny spotkały się w rozgrywkach AFC. Tym razem wygrał zespół z Baltimore i w 2013 roku więcej widzów obejrzało reklamy w tym mieście.

	Oglądalność Super Bowl 2012 (wygrywa Boston)	**Oglądalność Super Bowl 2013 (wygrywa Baltimore)**
Boston	56,7	48,0
Baltimore	47,9	59,6

Hal Varian, główny ekonomista Google'a, Michael D. Smith, ekonomista w Carnegie Mellon, i ja wykorzystaliśmy te dwa mecze oraz wszystkie inne finały Super Bowl pomiędzy 2004 i 2013 rokiem, by sprawdzić, czy reklamy nadawane w trakcie rozgrywek futbolowych są skuteczne, a jeżeli tak, to w jakiej mierze. Przyjrzeliśmy się zwłaszcza reklamom filmów nadawanym w czasie zawodów, bo chcieliśmy się przekonać, czy sprzedaż biletów do kin wzrasta w tych miastach, w których więcej osób ogląda spotkanie.

I rzeczywiście wzrasta. Mieszkańcy miast, skąd pochodzą drużyny grające o Super Bowl, wyraźnie chętniej chodzą na filmy reklamowane w czasie meczu niż mieszkańcy miast, które ostatecznie nie weszły do finału. W miastach finalistów więcej osób oglądało reklamy i więcej osób pofatygowało się do kin.

Alternatywne wyjaśnienie tego zjawiska może być takie, że jeżeli twoja drużyna weszła do finału, to masz większą ochotę na kino.

Zbadaliśmy jednak filmy o podobnym budżecie, które miały premiery mniej więcej w tym samym czasie, ale których nie reklamowano w trakcie rozgrywek. Na te filmy w miastach finalistów nie poszło jednak więcej widzów.

No więc jak się pewnie domyślaliście, reklama jest skuteczna. Nie ma w tym nic zaskakującego.

Ale reklama nie jest po prostu skuteczna. Jest niewiarygodnie skuteczna. Prawdę mówiąc, kiedy zobaczyliśmy wyniki, sprawdziliśmy je dwa, trzy, a nawet cztery razy, bo analizowane reklamy przyniosły tak duże zyski. Za reklamę filmu w czasie Super Bowl trzeba było przeciętnie zapłacić trzy miliony dolarów – lecz dochody ze zwiększonej sprzedaży biletów wynosiły średnio osiem milionów trzysta tysięcy, inwestycja zwracała się więc w proporcji 2,8 do 1.

Nasze wyniki potwierdziło dwóch ekonomistów, Wesley R. Hartmann i Daniel Klapper, którzy niezależnie od nas przeprowadzili podobne badania, dotyczące reklam piwa i napojów bezalkoholowych w czasie Super Bowl[16], i również powoływali się na zwiększoną oglądalność reklam w miastach drużyn występujących w finale. Reklamy emitowane podczas rozgrywek są niezwykle kosztowne, rezultaty naszych analiz wykazują jednak, że ich skuteczność tak bardzo zwiększa popyt na reklamowany produkt, że firmy właściwie płacą za nie zdecydowanie za mało.

A co to wszystko znaczy dla naszych przyjaciół z tamtej firmy elektronicznej, dla której pracował Levitt? Możliwe, że reklamy nadawane w czasie Super Bowl są najskuteczniejsze ze wszystkich, lecz z naszych badań wynika, że wzmożona reklama przed Dniem Ojca też wydaje się dobrym pomysłem.

Jedną z zalet „eksperymentu Super Bowl" jest to, że nie trzeba było nikogo wyznaczać do udziału w grupie roboczej czy kontrolnej. Test udało się przeprowadzić dzięki szczęśliwym zbiegom okoliczności w czasie meczu futbolowego. Dokonał się po prostu naturalnie – stąd mówimy tu o eksperymentach natury. Dlaczego miałaby to być zaleta? Bo nienaturalne, kontrolowane eksperymenty losowe, wprawdzie

niezwykle efektywne i łatwe do przeprowadzenia w dobie cyfryzacji, nie zawsze dają się zrealizować.

Czasami się z nimi spóźniamy. A czasami, jak w wypadku tej firmy elektronicznej, która nie chciała zbadać skuteczności swojej kampanii reklamowej, jesteśmy zbytnio zaangażowani emocjonalnie w odpowiedź, by ją zweryfikować.

Czasem jednak przeprowadzenie eksperymentu jest zwyczajnie niemożliwe. Przypuśćmy, że chcecie sprawdzić, co stanie się w danym kraju, kiedy straci przywódcę. Czy wywoła wojnę? Czy załamie się jego gospodarka? A może nic szczególnego się nie zmieni? Nie dowiemy się tego, bo nie możemy przecież po prostu pozabijać diabli wiedzą ilu prezydentów i premierów, żeby się przekonać, co będzie potem. Byłoby to zresztą nie tylko niemożliwe, ale też niemoralne. W ciągu ostatnich kilkudziesięciu lat amerykańskie uniwersytety powołały do życia tak zwane instytucjonalne komisje oceniające (Institutional Review Boards), które opiniują, czy proponowany eksperyment jest etyczny.

Co w takim razie zrobić, jeśli chcemy poznać skutki przyczynowe jakiegoś scenariusza, a kontrolowany eksperyment byłby nieetyczny albo niemożliwy do przeprowadzenia? Wtedy właśnie możemy wykorzystać to, co ekonomiści nazywają eksperymentem naturalnym, używając słowa „natura” w sensie na tyle szerokim, żeby obejmowało mecze futbolowe.

Trudno powiedzieć, czy to dobrze, czy źle (no dobrze: oczywiście, że źle), ale czynniki losowe odgrywają w naszym życiu wielką rolę. Nikt nie wie na pewno, kto czy co rządzi światem, jedno pozostaje wszak niewątpliwe: ktokolwiek lub cokolwiek to jest – prawa mechaniki kwantowej, Pan Bóg albo jakiś pryszczaty młodzieniec w bokserkach, tworzący w komputerze symulację wszechświata[17] – działania tych istot czy sił nie podlegają ocenie żadnej komisji.

Natura eksperymentuje z nami nieustannie. Oto dwie osoby zostają postrzelone. Jedna kula zatrzymuje się tuż przed uderzeniem w ważny organ życiowy, druga nie. Takie pechowe czy szczęśliwe zdarzenia sprawiają, że życie okazuje się niesprawiedliwe. Ale – chociaż to może niewielka pociecha – ułatwiają badania ekonomistom, którzy wykorzystują przypadkowość życia, aby wykrywać skutki przyczynowe.

Z czterdziestu trzech amerykańskich prezydentów szesnastu usiłowano zamordować w zamachach i czterech z nich zginęło[18]. To, że reszta przeżyła, było zasadniczo dziełem przypadku.

Weźmy Johna F. Kennedy'ego i Ronalda Reagana[19]. Obaj odnieśli niebezpieczne rany postrzałowe – pierwszego kula trafiła w głowę, a drugiego w okolice serca. W wypadku JFK pocisk eksplodował w mózgu, wkrótce potem powodując zgon – ale kula, która trafiła Reagana, zatrzymała się kilka centymetrów od serca, dzięki czemu lekarze zdołali uratować mu życie. Reagan przeżył, a JFK zmarł – przypadkiem, wskutek pecha.

Do podobnych zamachów dochodzi na całym świecie, a o tym, czy zagraniczni politycy giną, czy wychodzą z opresji cało, również decyduje przypadek. Weźmy Achmata Kadyrowa z Czeczenii i Adolfa Hitlera, kanclerza Niemiec. Życiu obu z nich zagrażały uzbrojone ładunki wybuchowe. Kadyrow padł ich ofiarą, a Hitler zmienił plany i na kilka minut przed wybuchem bomby wyszedł, by zdążyć na pociąg – dzięki temu przeżył. Kadyrow zginął na miejscu[20].

Spróbujmy zatem odwołać się do tej zimnej losowości natury – która zabiła Kennedy'ego, ale nie Reagana – żeby zobaczyć, co się na ogół dzieje, kiedy ginie przywódca jakiegoś kraju. Taką analizę przeprowadziło dwóch ekonomistów, Benjamin F. Jones i Benjamin A. Olken. Grupę kontrolną stanowią tutaj państwa, na których przywódców przeprowadzono prawie udane zamachy, czyli na przykład Stany Zjednoczone w połowie lat osiemdziesiątych, a grupę kontrolną kraje, gdzie zamachy na ważnych polityków się powiodły – czyli na przykład Stany Zjednoczone w połowie lat sześćdziesiątych.

A zatem co się dzieje, kiedy ktoś zamorduje przywódcę waszego kraju?[21] Jones i Olken twierdzą, że udane zamachy mają niezwykle poważny wpływ na bieg historii, a kraje, w których do nich doszło, wkraczają na zupełnie nowe ścieżki rozwoju. Nowy przywódca sprawia, że państwa dotąd spokojne wywołują wojny, a państwa wojownicze dążą do pokoju. Nowy przywódca wpędza kraje o świetnie prosperującej gospodarce w kryzys, a w krajach ekonomicznie słabych pobudza ich rozwój, zaczynają więc dobrze prosperować.

Można nawet powiedzieć, że rezultaty tego naturalnego eksperymentu, opartego na zamachach na życie ważnych polityków, obalają pokutujące przez kilkadziesiąt lat potoczne poglądy na temat sposobu funkcjonowania państwa. Wcześniej znaczna część ekonomistów skłaniała się do przekonania, że politycy to tylko marionetki, sterowane z zewnątrz przez takie czy inne siły. Badania Jonesa i Olkena temu zaprzeczają.

Wiele osób nie uznałoby takiego naturalnego eksperymentu za przykład wykorzystania big data. Liczba zamordowanych polityków, jak i takich, którzy cudem uszli z życiem, była na pewno nieduża, podobnie jak liczba wojen, jakie wybuchły albo nie wybuchły po zamachach. Z kolei ekonomiczne zbiory danych, niezbędne do tego, by określić trajektorię gospodarki jakiegoś kraju, bywały owszem, pokaźne, lecz w większości pochodziły z czasów poprzedzających cyfryzację.

Mimo wszystko takie naturalne eksperymenty, chociaż na razie wykorzystywane niemal wyłącznie przez ekonomistów, są skuteczne i będą zyskiwać na znaczeniu w epoce dającej nam dostęp do liczniejszych, lepszych i większych zbiorów danych. Uczeni innych dyscyplin nie będą mogli zbyt długo ignorować tego narzędzia wiedzy.

No i tak, teraz staje się już chyba jasne, że w rozwoju analizy danych pierwsze skrzypce grają ekonomiści. W każdym razie ja bardzo chciałbym w to wierzyć, gdyż studiowałem właśnie ekonomię.

Gdzie jeszcze możemy natrafić na eksperymenty naturalne – czyli sytuacje, w których to wypadki losowe dzielą ludzi na grupy robocze i kontrolne?

Najprostszym przykładem są wszelkiego rodzaju loterie, które ekonomiści zresztą uwielbiają – nie dlatego, żeby mieli w nie grać, hazard uważamy za irracjonalny – tylko aby je badać. Jeżeli z maszyny losującej wyskoczy piłeczka pingpongowa opatrzona trójką, to pan Jones będzie bogaty. Jeżeli szóstką, bogaty stanie się pan Johnson.

Aby sprawdzić skutki przyczynowe, występujące, kiedy pieniądze spadają nam nagle z nieba, ekonomiści zestawiają ze sobą loteryjnych zwycięzców i osoby, które kupiły losy, ale przegrały. Takie badania na

ogół wykazują, że wygrana na loterii nie czyni człowieka szczęśliwym na krótką metę[22], chociaż na długą, i owszem[17*].

Ekonomiści wykorzystują też loteryjną losowość, aby zbadać, jak zmienia się nasze życie, gdy sąsiad się wzbogaca. Analiza danych pokazuje, że jeśli sąsiad wygrywa na loterii, to może zmienić również i naszą przyszłość, bo na przykład częściej kupujemy wtedy drogie samochody w rodzaju bmw[23]. Dlaczego? Ekonomiści uważają, że prawie na pewno wynika to z zazdrości, która pojawia się, kiedy nasz bogatszy sąsiad kupuje nowe, luksusowe auto. Zachowanie to należy zatem przypisać ludzkiej naturze. Jeśli pan Johnson zobaczy pana Jonesa w nowiutkiej beemce, to zaraz będzie chciał mieć taką samą.

Niestety w wielu wypadkach nie stać go na taki samochód, i dlatego, jak uważają ekonomiści, sąsiedzi osób, które wygrały na loterii, dość często bankrutują[24]. W takiej sytuacji nie można po prostu beztrosko ścigać się z sąsiadami.

Eksperymenty naturalne wcale nie muszą być ewidentnie losowe, jak loterie. Kiedy zaczynamy szukać losowości, dostrzegamy ją wszędzie – i możemy ją stosować sami, by lepiej zrozumieć świat.

Częścią eksperymentów naturalnych są też lekarze. Co jakiś czas rząd USA, zasadniczo z arbitralnych powodów, zmienia zasady finansowania medyków opiekujących się pacjentami w ramach Medicare. W jednych hrabstwach wynagrodzenie lekarzy za niektóre badania i zabiegi wzrasta, w drugich spada.

Dwaj ekonomiści – Jeffrey Clemens i Joshua Gottlieb, mój kolega z ławy szkolnej – przeanalizowali skutki takich arbitralnych zmian. Czy lekarze zawsze zajmują się pacjentem tak samo i czy wykonują te badania oraz zabiegi, które naprawdę uważają za najbardziej konieczne? Czy może kieruje nimi motywacja finansowa?

Otóż analiza danych wykazuje wyraźnie, że lekarze często kierują się względami finansowymi[25]. W hrabstwach, gdzie wynagrodzenie jest wyższe, niektórzy z nich znacznie częściej zlecają wykonanie kosztowniejszych badań i zabiegów, takich jak operacje katarakty, kolonoskopie czy rezonans magnetyczny.

Pojawia się tu jednak ważkie pytanie: czy pacjenci mają się lepiej, kiedy korzystają z droższych zabiegów? Clemens i Gottlieb stwierdzili, że jeżeli w ogóle mają one wpływ na stan naszego zdrowia, to wyłącznie niewielki. Nie znaleźli też żadnych istotnych statystycznie zmian, jeśli chodzi o śmiertelność pacjentów. Z tego naturalnego eksperymentu wynikałoby więc, że kiedy damy lekarzom silniejszą motywację finansową do tego, aby ordynować podopiecznym niektóre droższe zabiegi i badania, znajdą się tacy, którzy rzeczywiście będą zlecać ich więcej, chociaż nie wpłynie to znacząco na poprawę stanu zdrowia pacjentów i raczej nie przedłuży im życia.

Eksperymenty naturalne mogą nam zatem pomóc rozstrzygać kwestie życia i śmierci, jak również takie problemy, które wielu młodym ludziom tylko wydają się śmiertelnie poważne.

Stuyvesant High School (liceum to bywa też nazywane w skrócie „Stuy") mieści się w dziesięciopiętrowym, wartym sto pięćdziesiąt milionów dolarów[26] ceglanym budynku z widokiem na rzekę Hudson, stojącym o kilka przecznic od World Trade Center na dolnym Manhattanie. Jednym słowem, ta szkoła robi wrażenie. Oferuje pięćdziesiąt pięć klas dla uczniów szczególnie uzdolnionych, naukę siedmiu języków i takie zajęcia fakultatywne, jak historia Żydów, literatura science fiction i literatura azjatycko-amerykańska[27]. Około dwudziestu pięciu absolwentów Stuya[28] trafia później na studia do jednego z uniwersytetów z Ivy League albo jakiegoś prestiżowego college'u. Tu pobierali nauki późniejsza profesor fizyki Lisa Randall, David Axelrod, strateg Baracka Obamy, nagrodzony Oscarem aktor Tim Robbins czy powieściopisarz Gary Shteyngart[29]. Na inauguracji roku szkolnego przemawiali w Stuyvesant High School między innymi Bill Clinton, Kofi Annan i Conan O'Brien[30].

Jest tylko jedna bardziej zdumiewająca rzecz niż absolwenci Stuya i jego edukacyjne propozycje dla uczniów, a mianowicie koszt nauki w tej szkole: zero dolarów. Bo jest to szkoła publiczna, zapewne najlepsza w kraju. W przeprowadzonej niedawno analizie, w której wykorzystano dwadzieścia siedem milionów recenzji autorstwa trzystu

tysięcy uczniów i rodziców, żeby sporządzić ranking wszystkich liceów publicznych w USA, Stuy zajął pierwsze miejsce[31]. Dlatego trudno się dziwić, że ambitni nowojorscy rodzice z klasy średniej i ich równie ambitne potomstwo dostają czasami fioła na punkcie tej szkoły.

Dla Ahmeda Yilmaza[18*], syna agenta ubezpieczeniowego i nauczycielki z Queensu, liceum imienia Stuyvesanta było „szkołą numer jeden".

„Rodziny robotnicze i imigranci traktują Stuya jako krok na drodze awansu społecznego – tłumaczy Yilmaz. – Gdy dziecko pójdzie do Stuya, to potem dostanie się na naprawdę porządny uniwersytet, z pierwszej dwudziestki. I wtedy cała rodzina będzie miała lepiej".

A jak można się dostać do Stuyvesant High School? Trzeba mieszkać w jednej z pięciu dzielnic Nowego Jorku i uzyskać odpowiednio dobry wynik na egzaminie wstępnym, to wszystko. Nie musicie mieć żadnych rekomendacji, nie pisze się wypracowań, nie przyjmuje się uczniów na podstawie pokrewieństwa z byłymi absolwentami (*legacy admissions*) i nie obowiązuje tu akcja afirmatywna, która polega na wyrównaniu szans kandydatów. Jeden dzień, jeden test, jeden wynik. Jeżeli zdobędziesz odpowiednią liczbę punktów, zostaniesz przyjęty.

Każdego roku w listopadzie blisko dwadzieścia siedem tysięcy młodych nowojorczyków podchodzi do egzaminu wstępnego do liceum. Rywalizacja jest zażarta, bo do Stuya dostaje się mniej niż pięć procent kandydatów[32].

Yilmaz mówi, że jego matka „harowała jak wół" i w przygotowanie syna do egzaminu włożyła wszystkie swoje skromne oszczędności. Po kilku miesiącach pracy – Yilmaz uczył się w każde popołudnie i w weekendy – chłopak uwierzył, że zda. Do dzisiaj pamięta dzień, gdy przyszła koperta z wynikami. Zabrakło mu odpowiedzi na dwa pytania.

Zapytałem go, jak się wtedy poczuł. „A jak się można czuć, kiedy rozsypuje ci się cały świat, a ty jesteś dopiero w szkole średniej?", odpowiedział.

Jego nagroda pocieszenia w żadnym razie nie była byle jaka – gdyż okazała się nią Bronx High School of Science, także ekskluzywna i wysoko oceniana w rankingach szkoła publiczna. Ale nie był to Stuy.

Poza tym Yilmaz uważał, że Bronx Science to szkoła raczej dla uczniów, którzy chcą się specjalizować w naukach politechnicznych. Po czterech latach nauki nie dostał się do Princeton, studiował w Tufts i próbował szczęścia w kilku zawodach. Dziś z jako takim powodzeniem pracuje w firmie technicznej, twierdzi jednak, że jego praca jest „ogłupiająca" i nie tak dobrze płatna, jak by sobie tego życzył.

Ponad dziesięć lat po egzaminach wstępnych do liceum Yilmaz przyznaje, że zastanawia się czasem, jak wyglądałoby jego życie, gdyby chodził do Stuya. „Wszystko byłoby inaczej – mówi. – Dosłownie, miałbym nawet innych znajomych". Zastanawia się, czy dzięki Stuyvesant High School uzyskałby lepsze wyniki na egzaminach SAT, dostałby się na uczelnię taką jak Princeton lub Harvard (obie uważa za znacznie lepsze niż Tufts) i może znalazłby lepiej płatną albo dającą mu więcej satysfakcji pracę.

Rozważania w rodzaju „co by było, gdyby" mogą stanowić dla ludzi zarówno rozrywkę, jak i zadawaną samemu sobie torturę. Jak wyglądałoby moje życie, gdybym zaczął emablować tę dziewczynę albo gdybym zaczęła podrywać tego chłopca? Gdybym podjął tę pracę? Gdybym trafił do tej szkoły? Na te pytania nie da się chyba jednak odpowiedzieć. Życie nie jest grą wideo. Nie można go odtwarzać wciąż na nowo wedle różnych scenariuszy, żeby uzyskać w końcu takie wyniki, jakich sobie życzymy.

Czeski pisarz Milan Kundera pisze lapidarnie na ten temat w swojej słynnej powieści *Nieznośna lekkość bytu*: „Człowiek nigdy nie może wiedzieć, czego ma chcieć, ponieważ dane mu jest tylko jedno życie i nie może go w żaden sposób porównać ze swymi poprzednimi życiami ani skorygować następnych"[19*].

Yilmaz nigdy nie będzie więc żył takim życiem, w którym udaje mu się uzyskać dwa punkty więcej na egzaminie wstępnym.

Ale może nam uda się zobaczyć, jak mogłoby wyglądać jego życie, jeśli przyjrzymy się losom dużej grupy absolwentów Stuyvesant High School?

Prymitywną, prostą metodą zbadania hipotetycznej przyszłości Yilmaza byłoby porównanie losów wszystkich uczniów, którzy poszli do

Stuya, i tych, którzy się tam nie dostali. Moglibyśmy przeanalizować, jakie osiągnęli wyniki na testach AP i SAT, i na jakich uczelniach później studiowali. Gdybyśmy to zrobili, przekonalibyśmy się, że absolwenci Stuyvesanta uzyskują znacznie lepsze rezultaty na testach badających wiedzę ogólną i zostają przyjęci na zdecydowanie lepsze uniwersytety. Jak jednak widzieliśmy już w tym rozdziale, takie informacje same w sobie nie są przekonujące. Może absolwenci Stuya odnoszą większe sukcesy po prostu dlatego, że ta szkoła w ogóle przyciąga lepszych uczniów. Korelacja nie dowodzi tu przyczynowości.

By sprawdzić przyczynowe skutki nauki w Stuyvesant High School, musimy porównać dwie niemal identyczne grupy badanych: tych, którzy uczyli się w Stuyu, i tych, którzy odpadli na etapie selekcji. Potrzebujemy eksperymentu naturalnego. Ale jak go skonstruować?

Odpowiedź brzmi: musimy poszukać uczniów takich jak Yilmaz, którzy byli bardzo, bardzo blisko zdania egzaminu do tego liceum[20*]. Uczniowie, którzy znaleźli się tuż pod kreską, to grupa kontrolna; uczniowie, którzy zdali egzamin, to grupa robocza.

Nie ma poważniejszych powodów, aby podejrzewać, że uczniowie znad i spod kreski znacząco od siebie odstają. Ostatecznie co sprawia, że ktoś dostaje o punkt albo o dwa więcej od kolegi? Może ten, kto uzyskał gorszy wynik, spał tamtego dnia dziesięć minut za krótko albo zjadł mniej pożywne śniadanie. A ten, kto uzyskał wynik lepszy, znał być może jakieś szczególnie trudne słowo występujące w teście, bo na przykład pojawiło się w rozmowie, którą przed trzema laty odbył ze swoją babcią.

Należy tutaj dodać, że ten rodzaj eksperymentu naturalnego, wykorzystującego znaczną liczbę odrzuconych kandydatów, jest tak skuteczny, że ma swoją nazwę wśród ekonomistów: nieciągłość regresyjna. Gdy można podzielić ludzi dokładnie na dwie grupy według jakiegoś kryterium liczbowego (nieciągłość), to można również przyjrzeć się losom osób, które znalazły się tuż pod i tuż ponad kreską podziału (regres).

Dwaj ekonomiści, M. Keith Chen i Jesse Shapiro, wykorzystali na przykład selekcję dokonywaną przez federalne zakłady karne, by sprawdzić, w jaki sposób ciężkie warunki więzienne wpływają na

przestępczość. Więźniowie federalni w Stanach Zjednoczonych podlegają punktacji, prowadzonej na podstawie kategorii popełnionych przestępstw i kryminalnej przeszłości. Ci, którzy uzyskają wysoki wynik, znajdą się w zakładzie karnym o podwyższonym rygorze, co oznacza rzadsze kontakty z ludźmi, większe ograniczenie swobody ruchów w więzieniu i najprawdopodobniej więcej przemocy ze strony strażników lub współwięźniów.

I znów: nie bylibyśmy obiektywni, porównując cały zbiór więźniów, którzy odsiadywali karę w więzieniach o zaostrzonym rygorze, z całym zbiorem więźniów odbywających wyrok w normalnych zakładach karnych. W więzieniach o zaostrzonym rygorze znajdziemy przecież znacznie więcej morderców i gwałcicieli, a w zwykłych zakładach karnych więcej drobnych złodziejaszków czy osób skazanych za przestępstwa związane z narkotykami.

Ci, którzy znaleźli się tuż pod albo nad kreską podziału, mieli jednak niemal identyczną przeszłość kryminalną i wywodzili się z podobnych środowisk. Ale jeden marny punkt różnicy oznaczał dla nich ekstremalnie różne konsekwencje penitencjarne.

Co się okazało? Otóż ekonomiści ustalili, że więźniowie odbywający karę w trudniejszych warunkach częściej wracali na drogę przestępstwa po wyjściu na wolność[33]. Zamiast odstraszać skazanych przed recydywą, trudne warunki hartowały ich i sprawiały, że po powrocie zza krat przejawiali jeszcze większą gwałtowność niż przedtem.

A co tego rodzaju eksperyment ujawnił w wypadku Stuyvesant High School? Badania przeprowadził zespół ekonomistów – Atila Abdulkadiroğlu, Joshua Angrist i Parag Pathak, pracownicy MIT oraz Duke University – którzy porównali osiągnięcia nowojorskich uczniów znad i spod kreski podziału na kandydatów przyjętych i odrzuconych. Przyjrzeli się setkom osób, którym tak jak Yilmazowi zabrakło punktu czy dwóch, aby uczyć się w Stuyu, a potem prześledzili także kariery szczęściarzy, którym tamtego dnia egzamin poszedł lepiej i którzy dzięki jednemu czy dwóm dodatkowym punktom dostali się do wymarzonej szkoły. Za miarę sukcesów uznano wyniki testów AP i SAT oraz miejsce rankingowe uczelni, na których badani ostatecznie studiowali.

Zdumiewające wnioski, do jakich doszli Abdulkadiroğlu, Angrist i Pathak, jasno określa tytuł artykułu, w którym je przedstawili: *The Ellite Illusion* [*Iluzja elity*]. Pożytki z nauki w Stuyvesant High School?[34] Żadne. Nic. Zero. Figa. Uczniowie spod i znad kreski uzyskiwali właściwie takie same wyniki w testach AP i SAT i uczęszczali potem na niemal równie prestiżowe uczelnie.

Jedyny powód, dla którego absolwenci Stuya osiągają w życiu więcej niż inni, jest zdaniem badaczy taki, że do egzaminów wstępnych do tej szkoły podchodzą generalnie uczniowie zdolniejsi. Sam Stuy nie sprawia jednak, że lepiej dają sobie radę na testach AP czy SAT albo dostają się na lepszy uniwersytet.

„Ostra rywalizacja na egzaminach wstępnych o miejsce w tej szkole wydaje się nie przekładać na późniejsze lepsze wyniki absolwentów w nauce", piszą członkowie zespołu.

Dlaczego więc najprawdopodobniej nie ma większego znaczenia, do jakiej szkoły uczęszczasz? Istnieją inne badania, które mogą nam pomóc uzyskać odpowiedź na to pytanie. Weźmy dwie uczennice z Nowego Jorku, Sarah Kaufmann i Jessicę Eng, które już od dzieciństwa marzyły, by dostać się do Stuya. Kaufmann ledwo zdała egzamin, jednym punktem. „Chyba już nigdy w życiu nie zaznam takich emocji", mówi. Natomiast wynik Eng plasował ją tuż pod kreską; zabrakło jednego punktu. Kaufmann poszła do upragnionej szkoły, Eng nie.

Jak potoczyło się ich życie później? Kariera zawodowa obu pań była udana i satysfakcjonująca – jak w wypadku wielu ludzi, którzy należą do pięciu procent nowojorczyków z najlepszymi wynikami w testach AP i SAT. Jak na ironię, Eng jednak lepiej wspomina czasy licealne niż Kaufmann. Bronx Science, szkoła, do której chodziła Jessica, to jedyne w USA liceum z własnym muzeum Holokaustu. Eng odkryła w sobie miłość do muzealnictwa i poszła na studia antropologiczne na Cornell University.

Natomiast Kaufmann czuła się trochę zagubiona w Stuyu, gdzie uczniowie nieustannie toczyli bezwzględną walkę o dobre stopnie, a poza tym uważała, że kładzie się tam za duży nacisk na testy zamiast na zajęcia lekcyjne. Okres nauki w Stuyvesant HS wspomina „ze

zdecydowanie mieszanymi uczuciami", chociaż czegoś się tam jednak nauczyła. Zrozumiała na przykład, że pójdzie na studia na uczelnię o profilu humanistycznym, gdzie przywiązuje się większą wagę do zajęć. Przyjęto ją na uniwersytet, o którym marzyła, Wesleyan. Odkryła tam w sobie pasję do pomagania ludziom i pracuje dzisiaj jako prawniczka, reprezentująca ludzi ubogich i pokrzywdzonych przez los.

Dostosowujemy się do warunków życiowych, a ci z nas, którym się wiedzie, potrafią dostrzec dobre strony każdej sytuacji. Czynnikami, które pomagają nam w odniesieniu sukcesu, są talent i motywacja. Nie jest natomiast ważne, kto wygłasza mowę inauguracyjną na rozpoczęcie roku szkolnego w naszym liceum, tak jak nie liczą się żadne inne atrakcje, jakie oferują najsłynniejsze szkoły.

Mówimy wprawdzie tylko o jednym badaniu, a jego rezultaty chyba osłabia dodatkowo fakt, że większość uczniów, którzy nie dostali się do Stuya, wylądowała w innych dobrych szkołach. Istnieje jednak coraz więcej dowodów, że chociaż do dobrej szkoły na pewno warto chodzić, to z nauki w najlepszym liceum Stanów Zjednoczonych płynie ostatecznie niewiele korzyści.

Spójrzmy także na uniwersytety. Czy to ważne, że uczęszczało się na jeden z najlepszych na świecie, jak Harvard, albo skończyło chociaż niezłą uczelnię, jak Penn State?

Istnieje tu wyraźna korelacja między rankingiem uniwersytetów a zarobkami. Po dziesięciu latach pracy zawodowej przeciętny absolwent Harvardu zarabia sto dwadzieścia trzy tysiące dolarów rocznie, a przeciętny absolwent Penn State mniej więcej osiemdziesiąt siedem tysięcy[35].

Lecz i ta korelacja wcale nie implikuje przyczynowości.

Stacy Dale i Alan B. Krueger, kolejni ekonomiści, wpadli na znakomity pomysł, jak sprawdzić, czy ukończenie elitarnego uniwersytetu wpływa na zarobki absolwentów. Dysponowali dużym zbiorem danych, zawierającym wiele informacji o licealistach, w tym na jakie uniwersytety składali papiery, gdzie zostali przyjęci, gdzie ukończyli studia, z jakich pochodzili rodzin i ile zarabiali już w dorosłym życiu.

By stworzyć grupę roboczą i kontrolną, Dale i Krueger zestawili ze sobą uczniów z podobnych rodzin, którzy zostali przyjęci na te same uczelnie, ale koniec końców wybrali inne. Byli tacy, których przyjęto na Harvard, ale którzy skończyli jednak Penn State – może dlatego, żeby być bliżej swojego chłopaka albo dziewczyny, lub też dlatego, że na Penn State wykładał jakiś profesor, do którego chcieli chodzić na zajęcia. Studenci ci, krótko mówiąc, byli tak samo utalentowani jak ci, którzy poszli na Harvard, ostatecznie kształcili się jednak gdzie indziej.

Co się więc działo, gdy dwaj studenci o podobnym pochodzeniu i statusie dostawali się na Harvard, ale jeden z nich wybierał Penn State? Wnioski okazały się tak samo zdumiewające jak w wypadku Stuyvesant High School. Otóż badani zarabiali po studiach mniej więcej tyle samo. Jeżeli pensja jest tu jakimś narzędziem pomiarowym, to równie zdolnym studentom, którzy wybierają różne uczelnie, finansowo powodzi się w życiu podobnie[36].

W gazetach roi się od artykułów o odnoszących oszałamiające sukcesy absolwentach uczelni z Ivy League; wymienia się wśród nich na przykład założyciela Microsoftu, Billa Gatesa, albo twórców Facebooka, Marka Zuckerberga i Dustina Moskovitza, którzy skończyli Harvard (chociaż należy przyznać, że wszyscy wymienieni przerwali naukę przed czasem i nie uzyskali dyplomów, co nasuwa dodatkowe pytania o wartość wykształcenia zdobywanego w uniwerkach z Ivy League).

Znamy też jednak opowieści o ludziach na tyle utalentowanych, że chociaż dostali się na jeden z nich, to wybierali mniej prestiżową uczelnię, a mimo to odnosili gigantyczne sukcesy; do takich osób należy Warren Buffett[37], który zaczynał w Wharton School na University of Pennsylvania, czyli w należącej do Ivy League szkole biznesu, przeniósł się jednak na University of Nebraska, do Lincoln, ponieważ ta uczelnia była tańsza, a Buffett nie znosił Filadelfii i uważał zajęcia w Wharton za nudne. Analiza danych wskazuje więc – przynajmniej jeśli jako kryterium przyjąć zarobki – że wybór mniej prestiżowego uniwersytetu to dobra decyzja, sądząc po przykładzie Buffetta i jemu podobnych.

Ta książka nosi tytuł *Wszyscy kłamią*. Chodzi mi głównie o to, że ludzie okłamują znajomych, przyjaciół oraz samych siebie i podają nieprawdę w sondażach po to, żeby wyglądać lepiej niż w rzeczywistości.

Z drugiej strony i świat okłamuje nas, podając nieprawdziwe lub mylące dane. Owszem, pokazuje wielką liczbę absolwentów Harvardu, którzy odnieśli sukces, lecz już o wiele mniej osób, które skończyły na przykład uniwersytet Penn State, zakładamy więc, że ukończenie Harvardu da nam olbrzymią przewagę.

Mądrze odczytując wyniki eksperymentów natury, możemy poprawnie zinterpretować dane, które podsuwa nam świat – i dzięki temu dowiedzieć się, co jest rzeczywiście użyteczne, a co nie.

Eksperymenty naturalne mają także związek z tematem poprzedniego rozdziału, bo często wymagają niejako filmowego zbliżenia – a to na grupy kontrolne i robocze, a to na miasta w doświadczeniu reklamowym w czasie Super Bowl, na hrabstwa w badaniach finansowych Medicare albo na uczniów znad i spod kreski w licealnym eksperymencie ze Stuyvesant High School. Takie uszczegółowienie, jak wskazaliśmy w poprzednim rozdziale, często wymaga dużych i kompletnych zbiorów danych, na szczęście coraz powszechniej dostępnych w cyfryzującym się świecie. Nigdy nie wiadomo jednak, kiedy natura postanowi zrealizować swoje eksperymenty, a nie wystarczy mały sondaż, aby zweryfikować ich wyniki. Potrzeba nam dużo danych, żeby się czegoś nauczyć na podstawie zjawisk naturalnych. Potrzeba nam big data.

W kontekście omawianych tu eksperymentów – zarówno kontrolowanych, jak i naturalnych – muszę dodać jeszcze jedną ważną uwagę. Znaczna część tej książki skupia się na próbie zrozumienia zasad rządzących światem – na tym, jaką cenę zapłacił Obama za rasizm, ilu Amerykanów naprawdę jest gejami i jakie kompleksy fizyczne dręczą mężczyzn, a jakie kobiety. Ale eksperymenty obu wspomnianych rodzajów mają też bardziej praktyczne zastosowanie, gdyż dzięki nim można podejmować lepsze decyzje, a także przekonać się, które działania są skuteczne, a które nie.

Przedsiębiorstwa mogą się dowiedzieć, jak zdobyć więcej klientów. Rządzący mogą się dowiedzieć, jak dzielić refundacje, aby możliwie najlepiej motywować lekarzy. Uczniowie mogą się dowiedzieć, które szkoły i uczelnie będą dla nich najbardziej wartościowe. Eksperymenty demonstrują potencjał zbiorów big data, który domysły, wiedzę potoczną i niejasne korelacje zastąpi czymś, co naprawdę pozwala na działanie skuteczne, ponieważ oparte na przyczynowości.

CZĘŚĆ TRZECIA

ZBIÓR BIG DATA: DLACZEGO NALEŻY OBCHODZIĆ SIĘ Z NIM OSTROŻNIE

ROZDZIAŁ SIÓDMY

BIG DATA, BIG SRATA? CZEGO NIE POTRAFI ZBIÓR BIG DATA

„Seth, Lawrence Summers chciałby się z Tobą spotkać", zawiadamiał mnie w e-mailu jeden z promotorów mojego doktoratu, Lawrence Katz, w taki poniekąd enigmatyczny sposób, bo nie napisał mi, dlaczego Summers interesuje się tym, co robię, choć dowiedziałem się później, że doskonale znał powód tego zaproszenia.

Usiadłem w poczekalni przed gabinetem Summersa. Po pewnym czasie były sekretarz skarbu Stanów Zjednoczonych, były rektor Harvardu i laureat kilku najważniejszych nagród ekonomicznych na świecie poprosił mnie do środka.

Zaczął spotkanie od lektury mojego artykułu o wpływie rasizmu na wyniki wyborów z udziałem Baracka Obamy, który wydrukowała mu jego sekretarka. Summers potrafi czytać bardzo szybko. W trakcie lektury wysuwa czasem raptownie język w prawo, a jego oczy błyskawicznie suną na boki i w dół. Kiedy czyta artykuł z dziedziny nauk społecznych, przypomina mi wielkiego pianistę, odgrywającego jakąś sonatę. Jest tak skoncentrowany, że najwyraźniej zapomina o bożym świecie. Tamten mój trzydziestostronicowy artykuł skończył w niecałe pięć minut.

– Dowodzisz, że wpisy ze słowem „czarnuch" w Google'u świadczą o rasizmie – powiedział. – Wydaje się to możliwe. Tak samo jak i to, że dzięki tym wyszukiwaniom można przewidzieć, gdzie Obama dostanie mniejsze poparcie niż Kerry. To interesujące. Ale czy naprawdę możemy porównywać Obamę i Kerry'ego?

– Politolodzy uznali, że obaj wyznają podobną ideologię – odparłem. – Poza tym brak korelacji pomiędzy rasizmem wyborców i wynikami głosowań w Kongresie. Moja analiza pozostaje wiarygodna nawet wówczas, kiedy uwzględnimy takie czynniki jak demografia, religijność i posiadanie broni.

Tak rozmawiają ze sobą ekonomiści. Ożywiłem się.

Summers zamilkł na chwilę i obrzucił mnie uważnym spojrzeniem. Odwrócił się w stronę telewizora, nastawionego na CNBC, po czym znów spojrzał na mnie, z powrotem na ekran i raz jeszcze na mnie.

– Okej, podoba mi się ten artykuł – oświadczył. – Nad czym jeszcze pracujesz?

Była to chyba najbardziej ekscytująca intelektualnie rozmowa w moim życiu. Przez następną godzinę rozmawialiśmy o stopach procentowych, inflacji, policji środowiskowej, przestępczości, biznesie i dobroczynności. Nie bez powodu wielu ludzi tak się zachwyca Summersem, kiedy go już pozna osobiście. Miałem szczęście rozmawiać w życiu z wieloma osobami o nieprzeciętnej inteligencji, ale Summers wydaje mi się najinteligentniejszy z nich wszystkich. Ma obsesję na punkcie różnych nowych pomysłów i idei, co chyba często wpędza go w kłopoty. Ze stanowiska rektora Harvardu musiał zrezygnować, kiedy zasugerował, że kobiety rzadziej uprawiają nauki ścisłe, bo mężczyźni mają przeciętnie wyższe IQ. Jeżeli jakaś koncepcja go zainteresuje, mówi o niej głośno, nawet jeśli obraża czyjeś uczucia.

Nasze spotkanie przedłużało się już o pół godziny. Rozmowa była pasjonująca, lecz wciąż nie miałem pojęcia, co ja robię w jego gabinecie, kiedy powinienem wyjść ani skąd miałbym wiedzieć, w którym momencie to zrobić. Wydało mi się nawet, że może Summers zapomniał, po co mnie zaprosił.

Potem jednak zadał mi pytanie za milion, a może za miliard dolarów.

– Myślisz, że dzięki analizie danych można przewidzieć wzrosty i spadki na giełdzie?

Aha. Nareszcie poznałem powód, dla którego wezwał mnie do siebie.

Naturalnie nie on pierwszy zadał mi to pytanie. Ojciec wspierał moje niekonwencjonalne zainteresowania badawcze na ogół bezwarunkowo,

zdarzyło mu się jednak poruszyć temat pieniędzy. „Rasizm, molestowanie dzieci, aborcja – powiedział. – A czy dzięki tej twojej analizie nie dałoby się zarobić trochę forsy?" Kwestię tę poruszali także przyjaciele i inni członkowie mojej rodziny, jak również moi współpracownicy i nieznajomi w sieci. Wszyscy chcą wiedzieć, czy wykorzystując wyszukiwania w Google'u albo inne elementy zbiorów big data, da się oskubać giełdy. Tym razem zapytał o to były sekretarz skarbu USA, sprawa była więc poważna.

A zatem czy big data pozwala przewidywać, dokąd zmierza giełda? Krótka odpowiedź brzmi: nie.

W poprzednich rozdziałach omawialiśmy cztery moce big data, ten rozdział poświęcony jest jednak ograniczeniom wielkich zbiorów danych; porozmawiamy o tym, czego nie można zrobić za ich pomocą, a czego czasem robić nie powinniśmy. Zacznę od opowieści, jak mnie i Summersowi nie udało się przechytrzyć giełdy.

W rozdziale trzecim zauważyliśmy, że nowe dane najczęściej przynoszą większe korzyści tam, gdzie nauka osiąga na razie słabe wyniki. Niestety, to prawda, że łatwiej dowiedzieć się czegoś nowego na temat rasizmu, molestowania dzieci czy aborcji, niż opracować nową metodę badawczą, dzięki której można się przekonać, jak radzi sobie gospodarka. Jest tak dlatego, że już wydaje się ogromne środki na poszukiwanie takich metod, które zapewniłyby ich użytkownikom choćby najlżejszą przewagę w trafnej ocenie wydajności ekonomii kraju. Rywalizacja w finansach jest zażarta. Już samo to działało na naszą niekorzyść.

Summers, który nie jest człowiekiem znanym z tego, że unosi się nad cudzą inteligencją, był pewien, że fundusze inwestycyjne mocno nas wyprzedzają. Ogromny szacunek, jaki dla nich miał, zrobił na mnie wrażenie, podobnie jak jego przekonanie, że zdążyły tymczasem zrealizować wiele moich pomysłów. Z dumą podzieliłem się z nim pewnym algorytmem, który opracowałem, aby pozyskiwać dokładniejsze dane z Google Trends. Stwierdził, że to sprytne, lecz kiedy zapytałem, czy Renaissance, ilościowy fundusz inwestycyjny, mógł wpaść na taki algorytm, zachichotał i odparł: „Oczywiście, że tak".

Ale trudności z dotrzymaniem kroku funduszom inwestycyjnym to nie jedyny fundamentalny problem, jaki napotkaliśmy z Summersem, próbując wykorzystywać nowe, duże zbiory danych do przewidywania ruchów akcji na giełdzie.

KLĄTWA WYMIAROWOŚCI

Przypuśćmy, że aby prognozować zmiany na giełdzie, usiłujesz znaleźć szczęśliwą monetę – jedną jedyną, którą wyznaczasz na podstawie starannych testów. Przyjmujesz następującą metodę: numerujesz tysiąc monet od jednego do tysiąca, a potem każdego ranka przez dwa lata rzucasz nimi, zapisujesz, czy wypadł orzeł, czy reszka, i sprawdzasz, czy indeks Standard & Poor's poszedł tego dnia w górę, czy może raczej w dół. Po upływie wyznaczonego czasu przeglądasz swoje dane i voilà! Coś znalazłeś. Okazuje się, że w siedemdziesięciu i trzech dziesiątych procent przypadków, gdy w rzucie monetą numer 391 wypadał orzeł, indeks S&P rósł. Związek pomiędzy tymi zjawiskami jest zatem statystycznie istotny, nawet bardzo. Znalazłeś szczęśliwą monetę!

Rzucaj więc dalej monetą numer 391 każdego ranka i kupuj akcje, kiedy wypadnie orzeł. Koniec z T-shirtami na wyprzedaży i z tanimi barowymi obiadkami. Moneta 391 to twoja przepustka do lepszego życia!

Chyba nie.

Bo stałeś się kolejną ofiarą jednego z najbardziej diabolicznych aspektów „klątwy wymiarowości". Może cię ona dopaść, gdy masz dużo zmiennych, czyli „wymiarów" (w tym wypadku chodzi o tysiąc monet), a nieszczególnie dużo obserwacji (w tym wypadku pięćset cztery dni robocze w ciągu dwóch lat). Jeden z tych wymiarów – tu moneta 391 – może okazać się „szczęśliwy". Zmniejsz liczbę zmiennych – rzucaj tylko stu monetami – a okaże się, że wytypować taką szczęśliwą monetę będzie znacznie trudniej. Zwiększ liczbę obserwacji – próbuj przewidywać zachowanie indeksu S&P przez dwadzieścia lat – a okaże się, że wytypować szczęśliwą monetę będzie równie trudno.

Klątwa wymiarowości to poważny problem dla big data, bo nowsze zbiory danych często dają nam wykładniczo więcej zmiennych niż źródła tradycyjne – danych dostarcza bowiem każdy zespół badawczy, każda kategoria tweetów, i tak dalej. Na wielu osobach, które twierdzą, że są w stanie przewidzieć zmiany giełdowe, używając do tego zbioru big data, ciąży po prostu ta klątwa, a wszystko, co udało im się osiągnąć, to tylko znaleźć ekwiwalent monety 391.

Weźmy na przykład zespół informatyków z Indiana University i Manchester University, którzy twierdzili, że potrafią przewidzieć zmiany na giełdzie na podstawie treści naszych tweetów[1]. Napisali specjalny algorytm, aby zakodować nastroje społeczne, określane w oparciu o tweety, i używali technik podobnych do analizy emocjonalnej, którą opisałem w rozdziale trzecim. Zakodowali nie jeden nastrój, lecz wiele – szczęście, złość, życzliwość i tym podobne. Okazało się, że przewaga tweetów wskazujących na spokój, na przykład „Jestem spokojny", sugeruje, że wskaźnik Dow Jones po sześciu dniach urośnie. Założono nawet fundusz inwestycyjny, który miał bazować na tym odkryciu.

W czym leży tutaj problem?

Przede wszystkim w tym, że nasi uczeni z Indiany i Manchesteru badali nazbyt wiele elementów. Jeżeli będzie ich odpowiednio dużo, to wskutek działania przypadku przynajmniej jeden z nich okaże się w końcu statystycznie istotny. Zbadali za wiele emocji. Poza tym robili to na jeden, dwa, trzy, a nawet siedem dni przed zmianami na giełdzie, na której chcieli przewidzieć ruchy akcji. A wszystkie zmienne wykorzystywano po to, by wyjaśnić ograniczone do zaledwie kilku miesięcy wzrosty i spadki indeksów.

Spokój w tweetach przez sześć dni nie jest uzasadnionym prognostykiem giełdowym. Spokój przez sześć dni to ekwiwalent naszej hipotetycznej monety 391, a oparty na badaniu tweetów fundusz inwestycyjny został zlikwidowany zaledwie po miesiącu, bo nie przynosił zysków[2].

Z klątwą wymiarowości borykają się nie tylko fundusze inwestycyjne, które próbują oceniać giełdę na podstawie tweetów, ale też liczni

uczeni, usiłujący znaleźć genetyczny klucz do istoty człowieczeństwa.

Dzięki Human Genome Project możemy dzisiaj zebrać i przeanalizować kompletną mapę ludzkiego DNA. Moc tego osiągnięcia wydawała się potencjalnie olbrzymia.

Może uda nam się znaleźć gen odpowiedzialny za schizofrenię. Może znajdziemy gen, który wywołuje chorobę Alzheimera, Parkinsona i stwardnienie zanikowe boczne. Może znajdziemy gen odpowiadający za – ojej – inteligencję. Tylko czy rzeczywiście istnieje taki gen, dzięki któremu moglibyśmy znacząco zwiększyć swój iloraz inteligencji? Czy istnieje gen czyniący człowieka geniuszem?

W 1998 roku Robert Plomin, wybitny genetyk behawioralny, ogłosił, że znalazł odpowiedź na te pytania. Zgromadził zbiór danych obejmujący DNA oraz wyniki badań IQ setek studentów, a potem porównał DNA „geniuszy" – osób z IQ 160 albo wyższym – z DNA studentów o przeciętnym ilorazie inteligencji.

I znalazł pomiędzy nimi uderzającą różnicę. Uwidoczniała się ona w maleńkim zakamarku chromosomu 6, w tajemniczym i potężnym genie, wykorzystywanym w metabolizmie mózgu. Plomin stwierdził, że jedna z wersji tego genu, o nazwie IGF2R, występuje u osób wybitnie inteligentnych dwukrotnie częściej.

Zastanówcie się, jak wielu kwestii etycznych dotknęło odkrycie Plomina. Czy rodzice powinni móc poddawać swoje dzieci badaniom na obecność tego genu? Czy powinni móc usunąć płód, jeśli okaże się, że jest nosicielem jego słabszego wariantu? Czy powinniśmy modyfikować ludzi genetycznie po to, aby zapewnić im wyższe IQ? Czy IGF2R jest w jakiś sposób skorelowany z rasą? Czy w ogóle chcemy poznać odpowiedź na to pytanie? Czy powinniśmy kontynuować badania nad powiązaniem genetyki z IQ?

Zanim do zajęcia się którymkolwiek z tych niewygodnych pytań poczuli się zmuszeni bioetycy, genetyków, w tym Plomina, zastanowiła jeszcze bardziej podstawowa kwestia. Czy wniosek płynący z jego badań był trafny? To znaczy, czy IGF2R naprawdę odpowiada za nasze IQ? I czy faktycznie jest prawdą, że geniusze są dwa razy częściej nosicielami określonego wariantu tego genu?

Nie. Kilka lat po przeprowadzeniu swoich pierwszych badań Plomin uzyskał dostęp do danych zawierających DNA i wyniki IQ innej grupy osób. Tym razem jednak IGF2R nie korelował z ilorazem inteligencji i Plomin, jak na prawdziwego uczonego przystało, wycofał się ze swoich poprzednich twierdzeń.

Jest to jednak pewna generalna prawidłowość w badaniach nad związkiem genetyki i IQ. Najpierw naukowcy donoszą, że znaleźli taki wariant genowy, który rzekomo odpowiada za wyższą inteligencję, a potem otrzymują nowe dane i odkrywają, że ich pierwotna hipoteza była mylna.

Na przykład niedawno zespół uczonych pod kierownictwem Christophera Chabrisa zbadał dwanaście ważnych twierdzeń dotyczących wariantów genów prawdopodobnie odpowiadających za inteligencję. Przeanalizowano dane dziesięciu tysięcy osób, lecz w żadnym z dwunastu przebadanych przypadków nie udało się potwierdzić korelacji między określonym genem i IQ[3].

Na czym polega problem? I tym razem chodzi o klątwę wymiarowości. Uczeni wiedzą już dzisiaj, że ludzkie genomy różnią się od siebie milionami szczegółów i że mamy po prostu za dużo genów do przebadania.

Podobnie jeśli przeanalizujesz odpowiednio dużo tweetów, by sprawdzić, czy są w jakiś sposób skorelowane z giełdą, to znajdziesz w końcu taki, który wykaże z nią związek, chociaż jedynie czysto przypadkowy.

Jak można przezwyciężyć klątwę wymiarowości? Trzeba podchodzić do swojej pracy z pokorą i nie zakochiwać się w wynikach, które musimy poddać dodatkowym sprawdzianom. Na przykład zanim postawicie swoje życiowe oszczędności na monetę 391, warto najpierw sprawdzić, jak będzie się zachowywać przez kolejne dwa lata. Specjaliści od nauk społecznych nazywają to „testem poza próbką”. A im więcej zmiennych analizujecie, tym musicie być pokorniejsi. Im więcej zmiennych, tym test poza próbką powinien być bardziej rygorystyczny. Należy jednak również zapisywać wszystkie testy, którym się je poddaje, bo wtedy można określić precyzyjnie, jakie istnieje prawdopodobieństwo, że eksperymentator pada ofiarą klątwy

wymiarowości, i jak bardzo sceptycznie powinien podchodzić do swoich wyników. W tym miejscu wracamy do Larry'ego Summersa i do mnie. Oto, jak próbowaliśmy przechytrzyć gospodarkę...

Pierwszym pomysłem Summersa było wykorzystanie wyszukiwań sieciowych, by przewidywać sprzedaż produktów przynoszących największe zyski, na przykład iPhone'ów, mogło to bowiem przybliżyć przyszłe wyniki giełdowe takich firm jak Apple. I okazało się, że istnieje korelacja między wyszukiwaniami iPhone'ów w Internecie oraz ich sprzedażą. Gdy ludzie często szukają smartfonów w sieci, można się założyć, że na rynku sprzedaje się ich dużo. Ta informacja była już jednak niejako wkalkulowana w cenę akcji Apple'a. Bez względu na to, czy korzystały z danych internetowych, czy z innego źródła informacji, fundusze inwestycyjne również doszły do wniosku, że gdy zwiększa się liczba wyszukiwań iPhone'ów, to urządzenia te będą się sprzedawały jeszcze lepiej.

Potem Summers wpadł na pomysł, by przewidywać przyszłe inwestycje w krajach rozwijających się. Jeżeli duża liczba inwestorów zacznie pompować pieniądze w takie państwa jak Brazylia czy Meksyk, to ceny akcji tamtejszych firm wkrótce niewątpliwie się podniosą. Może więc udałoby się nam przewidzieć wzrost inwestycji za pomocą wpisów w Google'u, na przykład takich jak „inwestuj w Meksyku" albo „możliwości inwestycyjne w Brazylii". Okazało się to ślepą uliczką. Dlaczego? Bo takich wyszukiwań było za mało i zamiast ujawniać znaczące prawidłowości, składały się tylko na jeden wielki internetowy chaos.

Spróbowaliśmy dalej szukać wpisów dotyczących konkretnych akcji. Może jeśli ludzie wpisują „Goog", to znaczy, że chcą kupić akcje Google'a. Takie wyszukiwania zdawały się sugerować, że papiery tej czy innej firmy będą często sprzedawane i kupowane. Nie można jednak było na tej podstawie przewidzieć, czy ich cena pójdzie w górę, czy spadnie. Stosując tę metodę, nie wiedzieliśmy też, choć przecież było to ważne, czy ludzie chcą akcje kupować, czy raczej sprzedawać.

Pewnego dnia podekscytowany przedstawiłem Summersowi nowy pomysł: wyszukiwania typu „kupię złoto" wydawały mi się kiedyś korelować z późniejszą zwyżką cen tego kruszcu. Summers

odpowiedział, że powinienem sprawdzić, czy nadal tak jest. Tym razem jednak nie stwierdziłem takiej korelacji, być może dlatego, że jakiś fundusz inwestycyjny spostrzegł już wcześniej związek między wyszukiwaniami o podobnej treści i cenami złota.

Ostatecznie więc po kilku miesiącach prób nie znaleźliśmy niczego pożytecznego. Niewątpliwie gdybyśmy szukali korelacji ze stanem giełdy i gospodarki w każdym z miliardów wyszukiwań w Google'u, to udałoby się nam znaleźć jakąś jedną, choćby tylko słabą, zmienną. Ale najprawdopodobniej byłaby to po prostu kolejna moneta 391.

PRZESADNY NACISK NA TO, CO MIERZALNE

W marcu 2012 roku Zoë Chance, profesorka marketingu w Yale, uniwersytecie mieszczącym się w centrum New Haven w stanie Connecticut, wyjęła ze swojej służbowej skrytki pocztowej mały biały pedometr[4]. Zamierzała zbadać, czy to urządzenie, mierzące liczbę kroków robionych w ciągu dnia i przyznające za nie użytkownikowi punkty, może nas zmotywować do większej aktywności fizycznej.

To, co stało się później, jak opowiadała na wykładzie w TEDx, to jeden z koszmarów big data. Chance dostała takiej obsesji pedometrycznej i tak kompulsywnie pragnęła zdobywać coraz więcej punktów, że zaczęła chodzić byle gdzie, z kuchni do salonu, do jadalni albo do piwnicy, i spacerowała nawet po swoim gabinecie na uczelni. Chodziła wcześnie rano, późnym wieczorem, prawie przez cały dzień – robiła dwadzieścia tysięcy kroków na dobę. Sprawdzała pedometr setki razy w ciągu dnia, a większość czasu, który pozostawał jej na kontakty międzyludzkie, spędzała online na rozmowach z innymi użytkownikami tych instrumentów pomiarowych, omawiając sposoby poprawienia wyników. Tak bardzo chciała uzyskiwać lepsze rezultaty, że założyła pedometr swojej dwuletniej córeczce, gdy tylko mała nauczyła się chodzić.

Chance pragnęła poprawić swoją punktację tak chorobliwie, że kompletnie straciła do siebie dystans i zapomniała w końcu, po co to

wszystko robi – a miała przecież zadbać o swoją aktywność fizyczną, nie o to, żeby córka przeszła za nią kilka kroków więcej. Nie ukończyła też badań nad pedometrem, który wyrzuciła, gdy pewnego dnia straciła równowagę i upadła, zmęczona, próbując nabić więcej kroków na licznik. Jest zawodowym analitykiem danych, ale ta przygoda głęboko ją odmieniła. „Zaczęłam mieć wątpliwości, czy dostęp do dodatkowych danych ma zawsze dobre skutki", mówi.

To ekstremalny przykład, który wskazuje jednak na potencjalne problemy, jakie napotykamy, gdy chcemy wykorzystywać dane w procesie decyzyjnym. Liczby mogą nam zawrócić w głowie[5]. Możemy dostać na ich punkcie fiksacji, a wtedy tracimy z oczu ważniejsze sprawy. Mówiąc z pewną przesadą, Zoë Chance straciła z oczu całe swoje życie.

Nawet mniej obsesyjna fascynacja liczbami może przynieść negatywne skutki. Weźmy nacisk, jaki w XXI wieku kładzie się w amerykańskich szkołach na testy – oraz na oceny nauczycieli, sporządzane na podstawie wyników osiąganych przez ich uczniów. Owszem, próba bardziej obiektywnego spojrzenia na to, co się dzieje w klasie, jest uzasadniona, dzieje się w niej jednak przecież wiele rzeczy, których nie można zbyt łatwo wyrazić liczbowo. Poza tym duża presja skłaniała znaczną część nauczycieli do tego, aby przygotowywać dzieci właściwie tylko do testów – a nawet jeszcze gorzej, bo jak dowodzą w swoim artykule Brian Jacob i Steven Levitt, byli wśród nich i tacy, na szczęście nieliczni, którzy po prostu fałszowali wyniki uczniowskich sprawdzianów[6].

Problem polega na tym, że nie zawsze obchodzą nas zjawiska i rzeczy mierzalne. Możemy zmierzyć, jak uczniowie radzą sobie z pytaniami wielokrotnego wyboru, ale nie możemy już z taką samą łatwością zbadać krytycznego myślenia, ciekawości świata albo stopnia personalnego rozwoju. Próby zwiększenia jakiegoś konkretnego, łatwo mierzalnego wyniku, na przykład szkolnego testu czy liczby kroków zrobionych w ciągu dnia, nie zawsze pomagają osiągnąć to, czego chcemy.

Nie ustając w wysiłkach, aby ulepszyć swój portal, Facebook też napotyka takie niebezpieczeństwo. Dysponuje całymi tonami danych,

mówiących, w jaki sposób ludzie z niego korzystają, łatwo więc zauważyć, czy podobała im się jakaś historia z tablicy, czy użytkownicy ją komentowali, klikali w nią albo dzielili się ze znajomymi. Jak twierdzi jednak analityk danych z Facebooka Alex Peysakhovich – z którym współpracowałem, pisząc na te tematy – żaden ze wspomnianych czynników nie daje nam wyczerpującej odpowiedzi na ważniejsze pytania: jakie wrażenie pozostawia po sobie korzystanie z portalu? Czy dana historyjka połączyła użytkownika ze znajomymi? Czy powiedziała mu coś nowego o świecie? Czy go rozbawiła?

Albo weźmy rewolucję w danych bejsbolowych z lat dziewięćdziesiątych. Zamiast polegać na staroświeckich skautach, wiele drużyn przy podejmowaniu decyzji personalnych zaczęło coraz częściej korzystać ze skomplikowanych statystyk. Łatwo było policzyć w ten sposób zagrania ofensywne i pitching, owszem, ale już nie fielding, niektóre zespoły nie doceniały więc wagi defensywy. Nate Silver w swojej książce *Sygnał i szum* szacuje, że drużyna Oakland A's, korzystająca z analizy danych i opisana w *Moneyball*, zaprzepaszczała w połowie lat dziewięćdziesiątych od ośmiu do dziesięciu zwycięstw rocznie z powodu słabej obrony.

A rozwiązaniem nie zawsze jest więcej big data. Aby ogromny zbiór danych działał optymalnie, trzeba podlać go specjalnym sosem: ludzką oceną i niewielkim sondażem, czymś, co moglibyśmy nazwać „small data". W rozmowie z Silverem ówczesny główny menedżer i bohater *Moneyball*, Billy Beane, powiedział, że po lekturze jego książki zaczął zwiększać budżet zespołu na skauting.

Żeby wypełnić luki w olbrzymiej puli danych, Facebook też musi korzystać ze staroświeckich metod, czyli po prostu pytać ludzi, co myślą na dany temat. Setki użytkowników, wyświetlając tablicę, dostają codziennie pytania na temat zamieszczanych tam postów, co oznacza, że automatycznie gromadzone dane facebookowe (lajki, kliknięcia i komentarze) wspomagane są mniejszymi zbiorami danych („Chcesz zobaczyć ten post na swojej tablicy? Dlaczego?"). Tak, nawet firmy, które odniosły równie spektakularny sukces jak Facebook, korzystają czasami ze źródła informacji często lekceważonego w tej książce: z małych sondaży i ankiet.

I właśnie ze względu na potrzebę pozyskiwania takich niewielkich zbiorów danych jako wsparcia dla podstaw struktury firmy – potężnego zbioru kliknięć, lajków i postów – analityczne zespoły Facebooka wyglądają inaczej, niż moglibyście przypuszczać. Facebook zatrudnia psychologów społecznych, antropologów i socjologów właśnie dlatego, żeby znaleźć to, czego nie da się wyrazić w liczbach.

Na martwe pola w big data zaczynają zwracać też uwagę niektórzy nauczyciele. Powstał rosnący narodowy ruch na rzecz wspierania sprawdzianów ogólnych niewielkimi zbiorami danych. Coraz częściej przeprowadzane są małe sondaże wśród uczniów i coraz częściej zdarzają się sondaże wśród rodziców, a także wizytacje nauczycielskie, polegające na tym, że w trakcie lekcji nauczyciela obserwują inni doświadczeni pedagodzy.

„Szkoły zaczynają zdawać sobie sprawę, że nie powinny skupiać się wyłącznie na wynikach testów", mówi Thomas Kane, profesor pedagogiki na Harvardzie[7]. Trwające trzy lata badania pracowników Bill & Melinda Gates Foundation podkreślają wartość zarówno dużych, jak i małych zbiorów danych w pedagogice. Uczeni sprawdzili, co jest najskuteczniejszym narzędziem oceny pracy nauczycieli: wyniki testów, ankiety uczniowskie czy wizytacje. Okazało się, że najlepiej działa połączenie wszystkich tych trzech metod. „Każda z nich wnosi do oceny coś wartościowego", napisali autorzy w konkluzji swojego raportu[8].

Nawiasem mówiąc, kiedy pojechałem na spotkanie z Jeffem Sederem na Florydę, przekonywałem się właśnie, że wiele działań na zbiorze big data wykorzystuje małe zbiory danych po to, aby wypełnić widniejące w nim luki. Seder, jak pamiętacie, jest tym wykształconym na Harvardzie znawcą koni wyścigowych, któremu na podstawie informacji zdobytych dzięki wielkiemu zbiorowi danych udało się przewidzieć sukces American Pharoah.

Seder najpierw pokazał mi wszystkie swoje pliki komputerowe i wzory matematyczne, a dopiero potem przyznał, że dysponował jeszcze jedną bronią – była nią Patty Murray.

Murray jest ponadprzeciętnie inteligentna, jak Jeff, i ma elitarne kompetencje – ukończyła Bryn Mawr. Ona również porzuciła Nowy

Jork, żeby zamieszkać na wsi. „Wolę konie niż ludzi”, mówi Patty, która jest jednak bardziej tradycyjna w swoim podejściu do oceny koni. Podobnie jak wielu ekspertów, bada je osobiście, przygląda się ich chodom, szuka blizn i zasinień i przepytuje właścicieli.

Następnie przystępuje do pracy z Sederem i już wspólnie wybierają te konie, którym zamierzają dać rekomendacje. A Murray potrafi wyczuć takie problemy, których nie wykryje analiza danych Sedera, chociaż pochodzą z najbardziej innowacyjnego i najważniejszego zbioru danych zgromadzonego dotąd na temat koni wyścigowych.

Przewiduję nadejście rewolucji, opartej na odkryciach i ustaleniach dokonywanych dzięki zbiorowi big data. Ale to nie znaczy, że w wypadku każdego problemu możemy po prostu odwoływać się do analizy danych czy lekceważyć jej aspekty etyczne. Poza tym big data wcale nie wyklucza innych dróg poznania, jakie wytyczał człowiek przez tysiąclecia. Wszystkie te sposoby uzupełniają się wzajemnie.

ROZDZIAŁ ÓSMY

WIĘCEJ DANYCH, WIĘCEJ PROBLEMÓW? CZEGO NIE POWINNIŚMY ROBIĆ

Czasami moc big data jest tak imponująca, że można się jej przestraszyć. Wywołuje nieuniknione pytania natury etycznej.

ZAGROŻENIE ZE STRONY WSZECHWŁADNYCH KORPORACJI

Niedawno troje znanych ekonomistów – Oded Netzer i Alain Lemaire, obaj z Columbia University, oraz Michal Herzenstein z University of Delaware – szukało sposobów, aby ustalić prawdopodobieństwo, że pożyczkobiorca spłaci kredyt[1]. Korzystali z danych z Prosper, portalu pożyczkowego. Potencjalni pożyczkobiorcy opisują na nim krótko, na co potrzebują pieniędzy i jakim sposobem zamierzają je spłacić, a potencjalni pożyczkodawcy podejmują decyzję, czy udzielić kredytu. Z wyliczeń badaczy wynika, że około trzynastu procent klientów nie oddaje długów[2].

Okazało się, że silnym prognostykiem sugerującym, czy potencjalni pożyczkobiorcy spłacą pożyczkę, czy nie, jest język, jakim się posługują. Prognostyk ten zachowuje swoją ważność, nawet biorąc pod uwagę inne istotne informacje, które mogli zdobyć pożyczkodawcy na temat przyszłych klientów, w tym zdolność kredytową i wysokość dochodów.

Poniżej prezentuję dziesięć określeń, których, jak stwierdzili uczeni, używają starający się o pożyczkę. Pięć z nich koreluje ze spłatą

pozytywnie, a pięć negatywnie. Innymi słowy, pięciu z nich używają ludzie, którym można ufać, pięciu zaś osoby niegodne zaufania. Spróbujcie zgadnąć, które to określenia.

Bóg	spłacę
obiecuję	absolwent
nie mam długów	dziękuję
minimalna kwota spłaty	po opodatkowaniu
niższa stopa procentowa	szpital

Można by pomyśleć, a w każdym razie przynajmniej mieć nadzieję, że sympatyczna i ewidentnie religijna osoba, która daje nam swoje słowo honoru, najprawdopodobniej spłaci pożyczkę – ale wcale tak nie jest. Przeciwnie, tacy ludzie, jak powiadają dane, są mniej wiarygodni od przeciętnego pożyczkobiorcy.

A oto określenia, które sygnalizują, że klient spłaci pożyczkę:

Określenia używane w aplikacjach o pożyczkę przez osoby, które najprawdopodobniej ją spłacą

nie mam długów	minimalna kwota spłaty
niższa stopa oprocentowania	absolwent
po opodatkowaniu	

Określenia używane w aplikacjach o pożyczkę przez osoby, które najprawdopodobniej jej nie spłacą

Bóg	dziękuję
obiecuję	szpital
spłacę	

Zanim omówimy etyczne implikacje tych badań, zastanówmy się z pomocą autorów nad tym, co ujawniają nam one na nasz temat. Jak powinniśmy rozumieć słowa i wyrażenia, pogrupowane w dwóch powyższych kategoriach?

Przyjrzyjmy się w pierwszej kolejności językowi sugerującemu, że ktoś raczej spłaci pożyczkę. Takie frazy jak „niższa stopa procentowa" czy „po opodatkowaniu" wskazują na niejakie finansowe obycie ze strony pożyczkobiorcy, więc może nie powinno dziwić, że sugerują osobę, która odda dług. A jeśli osoba taka wspomina również o swoich osiągnięciach, jak ukończenie studiów (absolwent), i pisze, że nie ma długów, to można przypuszczać, że chyba wywiąże się ze swoich zobowiązań.

W następnej kolejności spójrzmy natomiast na język sugerujący niewiarygodność pożyczkobiorcy. Otóż na ogół jest tak, że jeśli ktoś zapewnia, że odda wam pieniądze, to właśnie ich nie odda. Im solenniejsza obietnica, tym bardziej prawdopodobne, że zostanie złamana, a ten, kto pisze „Obiecuję spłacić pożyczkę, tak mi dopomóż Bóg", należy do najmniej wiarygodnych dłużników. Odwoływanie się do naszego współczucia i litości – na przykład informacja, że pożyczkobiorca potrzebuje pieniędzy, bo ktoś z jego rodziny leży w szpitalu – również oznacza, że raczej nas nie spłaci. Mało tego, już samo wspomnienie dowolnego członka rodziny – męża, żony, syna, córki, matki czy ojca – sygnalizuje, że nie możemy liczyć na zwrot pieniędzy. Kolejnym słowem będącym oznaką braku wiarygodności jest „wyjaśnię" – bo jeśli ktoś usiłuje nam wytłumaczyć, w jaki sposób zamierza spłacić pożyczkę, to najprawdopodobniej tego nie zrobi.

Autorzy nie stworzyli jednak żadnej teorii, która uzasadniałaby, dlaczego składanie podziękowań świadczy o tym, że dziękujący nie dotrzyma słowa.

Podsumowując: jeżeli ktoś przedstawia szczegółowy plan spłaty swojego długu i wymienia inne zobowiązania finansowe, z których wywiązał się w przeszłości, to według naukowców najprawdopodobniej odda zaciągnięty kredyt, natomiast składanie obietnic i apelowanie do współczucia pożyczkodawcy stanowią wyraźny znak, że pieniądze nie zostaną zwrócone. Bez względu na powody takiego stanu rzeczy – oraz

na to, co o naturze ludzkiej mówi fakt, że obietnica stanowi jasny sygnał, iż składający ją człowiek nie dotrzyma słowa – naukowcy stwierdzili, że opisana analiza przynosi wyjątkowo cenne informacje w prognozowaniu niewiarygodności pożyczkobiorców. Ci, którzy powołują się na Pana Boga, nie spłacają pożyczek dwa i dwie dziesiąte razy częściej, i jest to jeden z najwyższych wskaźników niewiarygodności w tym badaniu.

Ale jego autorzy uważają też, że wywołuje ono pytania natury etycznej. Mieliśmy do czynienia z badaniem akademickim, niektóre firmy kredytowe donoszą jednak, że rozpatrując wniosek o pożyczkę, korzystają z danych z sieci. Czy jest to do przyjęcia? Czy chcemy żyć w świecie, w którym przedsiębiorstwa finansowe analizują nasze wypowiedzi po to, by prognozować, czy spłacimy kredyt? Jest to co najmniej dziwne, ale chyba też i przerażające.

Już w niedalekiej przyszłości starający się o pożyczkę klient będzie chyba musiał się martwić nie tylko swoją historią kredytową, ale też aktywnością w Internecie. I możliwe, że jego zdolność kredytowa będzie oceniana na podstawie kryteriów, które sprawiają wrażenie absurdalnych – bo między innymi na przykład wedle tego, czy używa wyrażenia „dziękuję" albo wzywa imienia Pana Boga. Nie dość na tym – bo jak zostanie w takiej sytuacji potraktowana kobieta, która naprawdę chce pomóc leżącej w szpitalu siostrze i prawie na pewno spłaciłaby pożyczkę? Wydaje się okropieństwem karać ją tylko dlatego, że większość ludzi, którzy twierdzą w zgłoszeniu, że potrzebują pomocy, aby zapłacić za jakieś świadczenia medyczne, po prostu kłamie. Świat, w którym może dochodzić do takich sytuacji, zaczyna wyglądać koszmarnie, dystopijnie.

Pytanie etyczne: czy korporacje mają prawo oceniać naszą zdolność do korzystania z ich usług, opierając się na abstrakcyjnych, chociaż statystycznie uzasadnionych kryteriach predykcji, niepowiązanych bezpośrednio z tymi usługami?

Pozostawmy na chwilę świat finansów i spójrzmy na dalej idące konsekwencje, na przykład związane z rekrutacją do pracy. Rozpatrując kandydatury nowych pracowników, firmy coraz częściej przeglądają media społecznościowe. Nie budzi to może wątpliwości etycznych, jeśli

szukają dowodów na to, że ktoś źle wyrażał się o swoim poprzednim pracodawcy lub ujawniał tajemnice służbowe. Może nawet da się usprawiedliwić takie postępowanie, jeśli nie przyjmiemy do pracy kandydata, bo odkryliśmy, że jego posty na Facebooku albo Instagramie świadczą o nadużywaniu przez niego alkoholu. Ale co będzie, jeżeli pracodawca znajdzie jakąś niewinną informację, która z jego punktu widzenia koreluje jednak z czymś istotnym?

Badacze z Cambridge University i Microsoftu dali pięćdziesięciu ośmiu tysiącom użytkowników Facebooka do wypełnienia różne testy, badające osobowość i inteligencję. Stwierdzili, że lajki na Facebooku często skorelowane są z wynikami IQ, ekstrawertycznością i sumiennością[3]. Na przykład użytkownicy Facebooka, którzy lubią Mozarta, burze z piorunami i półkoliste frytki, mają na ogół wyższy iloraz inteligencji, a ci, którzy lubią motocykle Harleya-Davidsona, zespół muzyki country Lady Antebellum albo stronę *I Love Being a Mom*, niższy. Część takich korelacji może wynikać z klątwy wymiarowości – jeśli zbadamy dość dużą liczbę zjawisk, niektóre wykażą korelacje losowe. Przynajmniej część naszych zainteresowań ewidentnie może być jednak skorelowana z IQ.

Mimo to wydaje się niesprawiedliwe, aby inteligentny człowiek, który lubi harleye, nie mógł dostać pracy odpowiadającej jego kwalifikacjom, ponieważ nieświadomie zasygnalizował na Facebooku, że nie należy do szczególnie bystrych.

Prawdę mówiąc, nie jest to zupełnie nowy problem. Już od dawna jesteśmy oceniani na podstawie czynników, które nie są bezpośrednio związane z naszymi wynikami w pracy – na podstawie słabego lub mocnego uścisku dłoni czy schludności ubioru. Ale niebezpieczeństwo rewolucji informacyjnej tkwi w tym, że im więcej aspektów naszego życia poddamy kwantyfikacji, tym częściej takie pośrednie sądy będą stawać się bardziej enigmatyczne, a zarazem inwazyjne. Trafniejsze prognozowanie może prowadzić do bardziej perfidnej i nikczemnej dyskryminacji.

Dokładniejsze dane mogą skutkować też jej inną formą, którą ekonomiści nazywają cenową. Firmy często się zastanawiają, jaką wyznaczyć cenę za swoje towary czy usługi. Najchętniej

zaproponowałyby najwyższą, którą klient gotów jest zapłacić, bo wtedy osiągałyby najwyższe możliwe zyski ze sprzedaży.

Większość firm stawia zwykle na cenę w zasięgu każdego klienta. Czasem jednak producenci i usługodawcy zdają sobie sprawę, że członkowie określonej grupy społecznej mogą zapłacić więcej. Dlatego kina pobierają wyższe opłaty za bilety od widzów w średnim wieku, będących u szczytu możliwości zarobkowych, niż od studentów czy seniorów, i dlatego linie lotnicze często liczą sobie więcej za bilet kupiony w ostatniej chwili. Uprawiają dyskryminację cenową.

Big data może pozwolić producentom i usługodawcom orientować się o wiele lepiej, ile zechcą zapłacić ich klienci, i w ten sposób umożliwi im oszukiwanie niektórych grup społecznych. Firma Optimal Decision Group jako pierwsza stosowała analizę danych po to, by przewidzieć, ile jesteśmy gotowi wydać na ubezpieczenie. Jak to robiła? Jej pracownicy korzystali z metod, które omawialiśmy już w tej książce. Znajdowali wcześniejszych klientów, możliwie najbardziej podobnych do tych nowych, którzy dopiero chcieli się ubezpieczyć, i sprawdzali, jak wysoką cenę byliby gotowi zaakceptować. Czyli działali „metodą na sobowtóra". Jest to ciekawa metoda, jeśli pomaga nam przewidzieć, czy bejsbolista wróci do swojej dawnej świetnej formy. I metoda wspaniała, jeśli pomaga nam wyleczyć chorego. Ale jeżeli umożliwia jakiejś firmie wyciągnięcie od was ostatniego grosza? To już jest nie w porządku. Mój rozrzutny brat słusznie mógłby narzekać, gdyby zapłacił online więcej niż taki sknerus jak ja.

Sferą życia, w której możliwość wzięcia pod lupę klientów wydaje się też potencjalnie niebezpieczna, jest hazard. Duże kasyna korzystają z czegoś w rodzaju „metody na sobowtóra", by lepiej poznać swych graczy. Po co? Bo chcą z nich czerpać możliwie największe zyski – chcą wyciągać z was jeszcze więcej kasy i napchać sobie kieszenie.

Wygląda to tak: kasyno zakłada, że każdy hazardzista ma wykształcony tak zwany próg bólu. Jest to wysokość przegranej, która wystraszy gracza na tyle, że przez dłuższy czas nie będzie wracał do stołu. Przypuśćmy na przykład, że „próg bólu" Helen wynosi trzy tysiące dolarów. Oznacza to, że jeśli przegra tę kwotę, to stracicie klienta, być może na kilka tygodni lub miesięcy. Jeżeli przegra dwa tysiące

dziewięćset dziewięćdziesiąt dziewięć dolarów, nie będzie zadowolona – kto lubi przegrywać pieniądze? Ale nie przeżyje aż takiego rozczarowania, żeby nie wrócić do was następnego wieczoru.

Wyobraźcie sobie na chwilę, że zarządzacie kasynem. I że Helen przyszła do niego pograć na jednorękich bandytach. Jaki powinien być jej optymalny wynik? Oczywiście chcecie, żeby zbliżyła się jak najbardziej do „progu bólu", nie przekraczając go jednak. Chcecie, żeby przegrała dwa tysiące dziewięćset dziewięćdziesiąt dziewięć dolarów, wystarczająco dużo, żebyście wy sporo zarobili, ale nie tak wiele, aby szybko nie wróciła grać dalej.

Jak można to osiągnąć? No cóż, istnieją sposoby, by Helen przestała grać, kiedy wreszcie straci określoną kwotę. Możecie na przykład zaproponować jej wtedy darmowy posiłek. Jeśli wasza propozycja będzie dość kusząca, hazardzistka odejdzie od automatów, żeby coś zjeść.

Napotykamy tu wszakże jeden poważny problem. Skąd wiadomo, jaki jest „próg bólu" Helen? Ludzie mają przecież różne „progi bólu". Dla naszej bohaterki to trzy tysiące, ale dla Johna mogą być dwa, a dla Bena dwadzieścia sześć tysięcy. Jeżeli skłonisz Helen, żeby przestała grać, gdy pożegna się z dwoma tysiącami dolarów, zostawisz zyski na stole. Jeżeli będziesz czekać za długo, czyli zaczniesz działać dopiero, kiedy jej przegrana przekroczy trzy tysiące, to stracisz klientkę na dłużej. A przecież Helen chyba nie zechce podać ci wysokości swojego „progu bólu", chociażby dlatego, że sama może go nie znać.

Co zatem robić w takiej sytuacji? Jeśli dotarliście do tego miejsca w mojej książce, to pewnie się domyślacie: trzeba przeprowadzić analizę danych. Musicie dowiedzieć się jak najwięcej o swoich klientach, poznać ich wiek, płeć, kod pocztowy i zwyczaje hazardowe. I na ich podstawie – na podstawie wygranych, przegranych, wzlotów i upadków – wyznaczyć szacunkowo „próg bólu".

Dlatego trzeba zebrać wszystkie dostępne informacje o Helen i znaleźć hazardzistów podobnych do niej – można powiedzieć, że jej sobowtórów – potem zaś obliczyć, ile bólu mogą znieść te osoby; zapewne tyle, co Helen. W ten sposób postępuje na przykład kasyno

Harrah's, korzystające z usług firmy Terabyte, która magazynuje big data.

Scott Gnau, główny menedżer Terabyte, tłumaczy w doskonałej książce *Super Crunchers* autorstwa Iana Ayresa, jak zachowują się menedżerowie kasyn, widząc, że jeden z ich stałych klientów zbliża się do „progu bólu": „Podchodzą do niego i na przykład mówią: «Widać, że ma pan dzisiaj trudny dzień. Ale wiemy, że lubi pan tutejszy steakhouse. Zapraszamy więc pana wraz z małżonką na kolację do naszej restauracji»".

Może się wydawać, że to szczyt wielkoduszności: befsztyki, darmowy posiłek. Ale leży on tak naprawdę w interesie kasyna, które po prostu próbuje powstrzymać swoich klientów, zanim przegrają tak dużo, że już nigdy więcej do niego nie wrócą. Powiedzmy sobie wprost: zarząd kasyna używa wyrafinowanych metod analitycznych, aby na dłuższą metę wyciągnąć od graczy jak najwięcej forsy.

Istnieją zatem powody, żeby się obawiać, że coraz skuteczniejsze sposoby wykorzystywania danych z sieci dadzą kasynom, firmom ubezpieczeniowym, pożyczkodawcom i innym podmiotom gospodarczym zbyt dużą władzę nad ludźmi.

Z drugiej strony big data pozwala też klientom bronić się przed firmami, które naliczają zbyt wysokie ceny lub oferują niepełnowartościowy towar.

Jedna ważna broń w tej walce to portale takie jak Yelp, publikujące opinie o restauracjach i innych placówkach usługowych. Niedawne badania przeprowadzone przez Michaela Lucę, ekonomistę z Harvardu, pokazały, do jakiego stopnia firmy są uzależnione od recenzji z Yelpa[4]. Luca zestawił je z danymi na temat sprzedaży w stanie Waszyngton i okazało się, że jedna gwiazdka mniej na Yelpie powoduje spadek dochodów restauracji od pięciu do dziewięciu procent.

W zmaganiach z biznesem konsumentom pomagają także portale, na których można porównywać ceny, jak Kayak czy booking.com. Autorzy książki *Freakonomia*[21*] piszą, że kiedy powstał pierwszy portal internetowy podający ceny, jakich żądały różne firmy za terminowe ubezpieczenie na życie, ceny te szybko i radykalnie spadły. Jeżeli jakaś

firma ubezpieczeniowa żądała za dużo, to klienci dowiadywali się o tym i wybierali inną. Jakie przyniosło im to oszczędności? Miliard dolarów rocznie.

Dane internetowe mogą więc podpowiedzieć firmom, których klientów unikać i których mogą bezkarnie wykorzystywać, ale mogą też podpowiedzieć klientom, jakich firm się wystrzegać i kto próbuje ich naciągać. Big data na razie wspiera obie strony w walce korporacji z klientami. Musimy jednak pilnować, żeby ta walka była nadal prowadzona fair.

ZAGROŻENIE ZE STRONY WSZECHWŁADNEGO PAŃSTWA

Kiedy były chłopak Adriany Donato pojawił się u niej w domu na przyjęciu urodzinowym, zauważyła od razu, że ma zły humor. Wiedziała, że jest wściekły i że walczy z depresją. Przyjmując zaproszenie od dwudziestodwuletniego Jamesa Stonehama na wspólną przejażdżkę samochodem, Adriana, dwudziestojednoletnia studentka zoologii, nie wiedziała tylko jednego – że przez ostatnie trzy tygodnie szukał w sieci wskazówek, jak zabić człowieka, interesował się, jakie kary przewiduje prawo za morderstwo, a od czasu do czasu zbierał też internetowe informacje o swojej byłej dziewczynie.

Gdyby była świadoma tego wszystkiego, zapewne nie wsiadłaby z nim do samochodu. I może tamtego wieczoru nie zginęłaby od ciosów nożem.

W filmie *Raport mniejszości* jasnowidze współpracują z policją, by zapobiegać przestępstwom, powstrzymując ich potencjalnych sprawców, zanim złamią prawo. Czy policja powinna dysponować zbiorem big data, by dławić przestępstwa w zarodku? Czy Donato powinna była chociaż zostać powiadomiona o złowieszczych wyszukiwaniach Stonehama w Internecie? Czy policjanci powinni go byli przesłuchać?

Najpierw należy przyznać, że mamy coraz więcej dowodów na to, że wyszukiwania internetowe na temat działalności przestępczej rzeczywiście są z nią skorelowane. Christine Ma-Kellams, Flora Or, Ji Hyun Baek i Ichiro Kawachi wykazali też, że wyszukiwania w Google'u dotyczące samobójstw są silnie powiązane z ich liczbą w danym stanie[5]. Evan Soltas i ja dowiedliśmy z kolei, że islamofobiczne wpisy w sieci – w rodzaju „nienawidzę muzułmanów" albo „śmierć muzułmanom" – przekładają się na liczbę przestępstw z nienawiści wymierzonych w wyznawców islamu. Im więcej osób deklaruje w Internecie, że chce coś zrobić, tym więcej ludzi to zrobi.

Jak powinniśmy zatem reagować w takiej sytuacji? Jest na to jeden prosty i stosunkowo niekontrowersyjny pomysł: możemy wykorzystywać dane, żeby lepiej rozdysponować środki. Jeśli w jakimś mieście znacznie zwiększa się liczba wyszukiwań związanych z samobójstwami, to możemy uczulić na to mieszkańców. Władze miasta albo organizacje pozarządowe mogłyby na przykład podawać informacje w mediach, gdzie osoby gotowe targnąć się na własne życie mogą szukać pomocy. A jeśli w mieście nagle znacząco podnosi się liczba wpisów typu „śmierć muzułmanom", to policja postąpiłaby mądrze, zmieniając swoje zwyczaje patrolowe. Można by na przykład oddelegować więcej funkcjonariuszy do ochrony miejscowych meczetów.

Natomiast jednego powinniśmy się raczej wystrzegać: nie wolno ścigać nikogo, dopóki nie popełni przestępstwa. Przede wszystkim jest to po prostu naruszenie prywatności. Zachodzi ogromna etyczna różnica między danymi, którymi rozporządza państwo na temat tysięcy czy setek tysięcy swoich obywateli, i danymi, jakimi miałaby rozporządzać policja na temat konkretnego człowieka, tak jak zachodzi różnica między ochroną miejscowego meczetu a przeszukaniem czyjegoś domu albo między prewencyjnym upowszechnianiem informacji mających zapobiegać samobójstwom i zamknięciem kogoś w szpitalu psychiatrycznym wbrew jego woli.

Są jednak również analityczne, pozaetyczne powody, dla których należałoby zachować najwyższą ostrożność, korzystając z danych dotyczących konkretnego człowieka. Istnieje bowiem także poważna

różnica między próbą przewidzenia tego, jak zachowa się miasto, i jak zachowa się jednostka ludzka.

Wróćmy zatem na chwilę do samobójstwa. Każdego miesiąca w amerykańskim Google'u pojawia się około trzech i pół miliona wpisów związanych z autodestrukcją[6], przy czym większość z nich zdradza skłonności samobójcze autorów – są to wyszukiwania typu „nastrój samobójczy", „popełnić samobójstwo" albo „jak popełnić samobójstwo". Oznacza to, że miesięcznie na stu Amerykanów pojawia się więcej niż jeden wpis związany z samobójstwem, co przypomina mi o cytacie z Nietzschego: „Myśl o samobójstwie jest wielką pociechą; dzięki niej niejedną złą noc przepędza się dobrze"[22*]. Dane z Google'a świadczą o tym, że to prawda, bo ludzie bardzo często o samobójstwie myślą. A jednak w Stanach Zjednoczonych popełnia się mniej niż cztery tysiące samobójstw miesięcznie. Myśl o odebraniu sobie życia jest szeroko rozpowszechniona, lecz sam akt już nie, nie miałoby więc chyba większego sensu, gdyby gliniarze pukali do drzwi każdego, kto chociaż raz wspomniał w Internecie o tym, że zastanawia się, czy nie palnąć sobie w łeb – już choćby z tej przyczyny, że wtedy policjanci nie mieliby czasu na żadne inne czynności operacyjne.

Albo weźmy owe niezwykle agresywne wyszukiwania islamofobiczne[7]. W 2015 roku w amerykańskiej sieci znalazło się około dwunastu tysięcy wpisów o treści „śmierć muzułmanom", a popełnionych na muzułmanach morderstw, zakwalifikowanych jako zbrodnie z nienawiści, odnotowano dwanaście. Najwyraźniej zatem ogromna większość ludzi będących autorami przerażających wpisów nie posuwa się do popełnienia czynu, do którego nawołuje innych.

Potrzebujemy trochę matematyki, żeby wyjaśnić różnicę między prognozowaniem zachowania konkretnego człowieka a prognozowaniem zachowania zbiorowości miejskiej. Oto prosty eksperyment myślowy: przypuśćmy, że mamy w mieście jeden meczet i milion mieszkańców. Załóżmy, że jeśli żaden z nich nie szuka w sieci wpisów „śmierć muzułmanom", to istnieje tylko jedna szansa na milion, że ktoś napadnie na muzułmańską świątynię; jeżeli natomiast ktoś szuka takich wpisów, to prawdopodobieństwo ataku gwałtownie rośnie i jego szansa wynosi już jeden do dziesięciu tysięcy. Przypuśćmy dalej,

że islamofobia znacznie się nasiliła i liczba wpisów w rodzaju „śmierć muzułmanom" zwiększyła się, powiedzmy, ze stu do tysiąca. Wtedy szanse, że meczet zostanie zaatakowany, wzrastają pięciokrotnie, z dwóch do dziesięciu procent, chociaż prawdopodobieństwo, że osoba, która wpisywała w wyszukiwarkę hasła w rodzaju „śmierć muzułmanom", faktycznie napadnie na meczet, wynosi nadal jeden do dziesięciu tysięcy.

Odpowiednią reakcją w takiej sytuacji nie jest wsadzenie do więzienia wszystkich, którzy szukali islamofobicznych wpisów, ani wizyta w ich domach. Istnieje jedynie mikroskopijna szansa, że któraś z tych osób popełni jakieś przestępstwo. Właściwą reakcją byłaby natomiast ochrona meczetu, istnieje już bowiem dziesięcioprocentowe prawdopodobieństwo, że zostanie on zaatakowany.

Jak z tego wynika, wiele budzących grozę wpisów nigdy nie prowadzi do krwawego działania.

To prawda, jest jednak przynajmniej teoretycznie możliwe, że niektóre typy wpisów i wyszukiwań mogą z całkiem dużym prawdopodobieństwem skończyć się makabrycznie. I jest również teoretycznie możliwe, że analitycy danych będą umieli w przyszłości stworzyć algorytm, dzięki któremu dałoby się stwierdzić, że wpisy Stonehama związane z jego byłą dziewczyną stanowiły poważny powód do obaw.

W 2014 roku mieliśmy około sześciu tysięcy wyszukiwań typu „jak zabić swoją dziewczynę" i czterysta morderstw, które popełnili partnerzy ofiar. Jeśli wszyscy mordercy dokonywali wcześniej takich wpisów w Internecie, to swoje zamiary spełniła jedna na piętnaście osób szukających linków do tej frazy. Oczywiście wielu, a nawet większość facetów, którzy zamordowali swoje partnerki, nie wpisywała takich treści w wyszukiwarkę. Oznaczałoby to, że realne prawdopodobieństwo, iż tego rodzaju wpisy prowadzą do zabójstwa, jest niższe, i to zapewne znacznie.

Jeśli jednak analitycy zdołaliby stworzyć algorytm, na podstawie którego dałoby się stwierdzić, że istnieje zagrożenie życia konkretnej osoby i wynosi ono, powiedzmy, jeden do stu, to chyba chcielibyśmy jakoś taką informację wykorzystać. W każdym razie osoba zagrożona

powinna chyba mieć prawo dostępu do informacji, że istnieje jedna szansa na sto, iż ktoś ją zamorduje.

Generalnie jednak musimy być bardzo ostrożni, używając danych z wyszukiwarek do prognozowania przestępstw na poziomie indywidualnym. Dane te mówią wyraźnie, że wielka, a nawet bardzo wielka liczba przerażających wpisów rzadko prowadzi do aktów przemocy czy zbrodni. A jak na razie brak jakichkolwiek dowodów, że organa ścigania byłyby zdolne z dużym prawdopodobieństwem przewidzieć taki czy inny krwawy czyn wyłącznie na podstawie analizy wpisów w sieci. Dlatego powinniśmy zachować daleko idącą ostrożność i nie pozwolić rządowi na prowadzenie działań przeciwko jednostce wyłącznie na podstawie danych z Internetu. Nie chodzi tu jedynie o powody etyczne czy logiczne, albowiem przynajmniej na razie przeciwko takiemu postępowaniu przemawiają również argumenty natury naukowej.

ZAKOŃCZENIE
ILU LUDZI CZYTA KSIĄŻKI DO KOŃCA?

Po podpisaniu umowy miałem już jasno sprecyzowane poglądy, jak powinienem skonstruować *Wszyscy kłamią*. Pamiętacie może, że na początku opisałem scenę rozgrywającą się w moim rodzinnym domu, przy stole, w Święto Dziękczynienia. Członkowie mojej rodziny zastanawiali się, czy jestem przy zdrowych zmysłach, i zadawali sobie pytanie, dlaczego, chociaż mam trzydzieści trzy lata, nie znalazłem jeszcze odpowiedniej dziewczyny.

Zakończenie było więc właściwie gotowe. Powinienem spotkać kobietę swojego życia i ożenić się z nią. I byłoby najlepiej, gdybym użył zbioru big data, żeby ją poznać. Może udałoby mi się też wpleść w tekst co ciekawsze szczegóły z okresu zalotów, a wówczas cała historia zostałaby zamknięta w ostatnim rozdziale, będącym listem miłosnym do mojej świeżo poślubionej żony, w którym opisałbym też dzień naszego ślubu.

Przykro mi, ale życie nie dopasowało się do takiej wizji. Pisząc *Wszyscy kłamią*, zamknąłem się w mieszkaniu i unikałem kontaktów ze światem, co chyba nie wpłynęło korzystnie na moje życie uczuciowe. Niestety wciąż szukam więc żony, ale co ważniejsze, potrzebowałem innego zakończenia.

Ślęczałem nad wieloma swoimi ulubionymi książkami, żeby zobaczyć, na czym polega dobry finał. I doszedłem do wniosku, że najlepsze z nich wydobywają na światło dzienne jakieś ważne spostrzeżenie, które wcześniej przez cały czas unosiło się tuż pod powierzchnią tekstu. Jeśli chodzi o tekst *Wszyscy kłamią*, to najistotniejsze autorskie przesłanie brzmi tak: nauki społeczne stają się

naukami ścisłymi. A te nowe, prawdziwe nauki społeczne mają na celu polepszenie warunków ludzkiego życia.

Na początku drugiej części książki omówiłem krytykę poglądów Freuda, której dokonał Karl Popper. Zauważyłem, że jego zdaniem obłąkańcza wizja świata wiedeńskiego uczonego nie jest naukowa. Ale nie wspomniałem o jednej rzeczy w związku z krytyką Poppera, która była czymś znacznie szerszym niż tylko atakiem na ojca psychoanalizy. Otóż Popper uważał, że żadna nauka społeczna nie jest w istocie nauką; jego zdaniem to, co robili tak zwani naukowcy z dziedzin społecznych, miało po prostu nieścisły i nieweryfikowalny charakter.

Co stało za jego krucjatą?[1] Popper znał najwybitniejszych intelektualistów swoich czasów – najlepszych fizyków, historyków i psychologów – dostrzegał jednak między nimi uderzającą różnicę. Słuchał fizyków i ufał ich dokonaniom. Jasne, czasem popełniali błędy. Jasne, czasem na manowce sprowadzał ich podświadomy subiektywizm. Fizycy ewidentnie uczestniczyli jednak w procesie poznawczym, umożliwiającym dotarcie do najbardziej podstawowych prawd fizycznych, które zwieńczyła teoria względności Einsteina. Natomiast kiedy rozmawiał z najsławniejszymi socjologami i przedstawicielami innych nauk społecznych, odnosił wrażenie, że słucha steku bredni.

Rzecz jasna, nie tylko Popper widział tę różnicę. Zapewne wszyscy się zgodzimy, że naukowcami z prawdziwego zdarzenia są fizycy, biolodzy i chemicy, którzy stosują starannie zaplanowane eksperymenty, żeby poznać świat. A wielu z nas uważa, że ekonomiści, socjologowie i psychologowie to pseudouczeni, którzy posługują się bezsensownym żargonem, aby zajmować profesorskie stanowiska na uczelniach.

Jeżeli nawet naprawdę tak kiedyś było, to rewolucja big data zmieniła ten stan rzeczy. Gdyby Popper jeszcze żył i mógł pójść na wykład Raja Chetty'ego, Jessego Shapiro, Esther Duflo czy (chcę zrobić sobie dobrze) mój, to mam podejrzenie, graniczące z pewnością, że zmieniłby zdanie. Szczerze mówiąc, myślę, że może nawet raczej zacząłby się zastanawiać, czy wielcy współcześni autorzy teorii strun naprawdę są rzetelnymi naukowcami, czy może jednak to zwyczajni fantaści, zajmujący się gimnastyką umysłową wyłącznie na własny użytek.

Czy przestępczość wzrasta, kiedy do kin trafiają filmy akcji? Czy jeśli reklama dociera do szerszego kręgu odbiorców, to więcej osób zaczyna używać reklamowanego produktu? Czy jeśli drużyna bejsbolowa odnosi zwycięstwa, kiedy chłopak ma dwadzieścia lat, to ten sam chłopak będzie jej wciąż kibicował jako czterdziestolatek? Są to jasne pytania, na które można udzielić jasnych odpowiedzi typu „tak" lub „nie". I można je znaleźć w oceanach niezafałszowanych danych.

W ten sposób uprawia się naukę, nie pseudonaukę.

Nie oznacza to jednak, że rewolucja w naukach społecznych objawi się nam w postaci prostych, ponadczasowych praw.

Marvin Minsky, nieżyjący już uczony z Massachusetts Institute of Technology, jeden z pierwszych na świecie zajmujących się sztuczną inteligencją, twierdził, że psychologia zeszła z właściwej jej ścieżki, ponieważ starała się naśladować fizykę, która odniosła sukces, odkrywając proste prawa, obowiązujące zawsze i wszędzie.

Tymczasem zdaniem Minsky'ego takim prawom ludzki umysł raczej nie podlega. Bardziej przypomina skomplikowany układ programów komputerowych, w którym jeden element koryguje pomyłki pozostałych. Nasza ekonomia i system polityczny są chyba podobnie złożone.

Dlatego rewolucja w naukach społecznych raczej nie będzie się wyrażać w postaci schludnych formułek typu $E = mc^2$. A jeśli ktoś wam mówi, że stanie się inaczej, to powinniście podchodzić do takich opinii sceptycznie.

Rewolucja zapewne okaże się stopniowa, a ukształtują ją kolejne badania i odkrycia. Powoli jednak zaczniemy lepiej rozumieć wieloaspektowy system ludzkiego umysłu i struktury społecznej.

Odpowiednie zakończenie jest podsumowaniem, lecz wskazuje także drogę ku temu, co nastąpi.

W wypadku tej książki to łatwe. Zbiory danych, które tu omawiałem, są rewolucyjne, ale dopiero zaczęto je eksploatować, pozostało więc jeszcze mnóstwo do odkrycia. Prawdę mówiąc, przygniatająca większość uczonych akademickich zlekceważyła eksplozję danych

wywołaną nadejściem epoki cyfrowej. Najsławniejsi seksuolodzy nadal trzymają się starych, wypróbowanych metod. Wypytują kilkuset ankietowanych o ich fantazje i pragnienia, ale nie biorą pod lupę takich portali jak PornHub w poszukiwaniu danych na temat tych osób. Najwybitniejsi językoznawcy analizują pojedyncze teksty, lecz w znacznej mierze lekceważą prawidłowości, które przejawiają się w miliardach książek. Metodologia takich nauk jak psychologia, politologia czy socjologia, której uczą się studenci, pozostała w większej części nietknięta przez rewolucję cyfrową. Olbrzymie, prawie całkowicie nieznane terytorium, otwarte dzięki proliferacji danych, pozostawiono tylko niewielkiej liczbie przyszłościowo myślących profesorów, buntowniczych studentów i hobbistów.

Ale to się zmieni.

Na każdą nową myśl, którą przedstawiłem w niniejszej książce, przypada sto równie ważnych myśli, jakie należy dopiero przeanalizować. Omawiane tu przeze mnie badania to zaledwie czubek czubka góry lodowej, rysa na rysie na powierzchni zjawisk.

A co nas czeka dalej?

Między innymi radykalna ekspansja pewnej metody, którą zastosowano w jednym z najskuteczniejszych badań medycznych wszech czasów. Otóż w połowie XIX wieku John Snow, angielski lekarz, postanowił znaleźć źródło epidemii cholery, panoszącej się w ówczesnym Londynie[2].

Wpadł na genialny pomysł: nanosił na mapę każdy odnotowany przypadek choroby. Okazało się, że cholera występuje najgęściej w pobliżu pewnej studni, co wskazywało na to, że rozprzestrzenia się wraz z zainfekowaną wodą, zaprzeczając rozpowszechnionemu wówczas przekonaniu, że epidemii sprzyja zanieczyszczone powietrze.

Zbiór big data umożliwia uszczegółowienia takich badań i ułatwia ich przeprowadzenie. Możemy przeanalizować dane z wyszukiwarki Google'a albo wszystkie inne dane cyfrowe dla każdej jednostki chorobowej. I możemy się przekonać, czy istnieją gdzieś na świecie enklawy, w których choroba ta występuje szczególnie często lub szczególnie rzadko, a potem sprawdzić, co takie miejsca mają ze sobą

wspólnego. Jakiś składnik w powietrzu? A może właśnie w wodzie? Jakieś normy społeczne?

Podobne analizy możemy przeprowadzić, szukając przyczyn migreny. Powodów powstawania kamieni nerkowych. Stanów lękowych, depresji, choroby Alzheimera, raka trzustki, nadciśnienia, bólu pleców, zaparć i krwotoków z nosa. Możemy przeprowadzić podobne analizy, szukając przyczyny dowolnego zjawiska. Analizę, jaką Snow mógł wykonać tylko jednokrotnie, możemy powtórzyć dzisiaj choćby czterysta razy (zaczynam już nad tym pracować, jeszcze kończąc tę książkę).

Taką prostą metodę i wykorzystanie big data do przeprowadzenia kilkuset analiz w krótkim czasie moglibyśmy nazwać nauką wyskalowaną (*at scale*). Tak, nauki społeczne i behawioralne z całą pewnością zostaną wyskalowane, uszczegółowione. Pomoże im w tym na przykład bliższe spojrzenie na sytuację zdrowotną społeczeństwa, a także testy typu A/B. Omawialiśmy je w kontekście firm, które chciały nakłonić klientów, żeby klikali w nagłówki i reklamy – i ta metoda była jak dotąd stosowana w tym celu. Ale za pomocą testów A/B można dokonywać ważniejszych i społecznie bardziej wartościowych odkryć niż to, że strzałka dodawana do reklam skłania internautów do zapoznania się z ich treścią.

Na przykład ekonomista z Northwestern University Benjamin F. Jones próbuje stosować testy A/B, by pomóc dzieciom w nauce[3]. Jones jest współtwórcą platformy EDU STAR, na której szkoły mogą losowo testować różne programy zajęć.

Oprogramowaniem dla oświaty zajmuje się wiele firm komputerowych. Na EDU STAR uczniowie po zalogowaniu uzyskują swobodny dostęp do różnych programów lekcyjnych. Następnie wypełniają krótkie testy, by sprawdzić, jak dobrze opanowali materiał. Dzięki temu szkoły mogą się dowiedzieć, które oprogramowanie najlepiej pomaga dzieciom w nauce.

Jak wszystkie wielkie platformy testowe, EDU STAR już odnotowuje zaskakujące wyniki. Wielu pedagogów z ogromnym entuzjazmem podchodziło na przykład do programu, który wykorzystywał gry komputerowe do nauki ułamków. No oczywiście, bo jeśli

przekształcimy matematykę w zabawę, uczniowie będą mieli z niej więcej frajdy, więcej się nauczą i osiągną lepsze rezultaty na testach, prawda? Nie. Dzieci uczące się ułamków z wykorzystaniem gier komputerowych osiągały na testach gorsze oceny niż dzieci kształcone w tradycyjny sposób.

Próba skłonienia dzieciaków, żeby się więcej uczyły, to jeden z bardziej ekscytujących i społecznie wartościowych sposobów użycia testów A/B, które Dolina Krzemowa pierwotnie stosowała po to, by internauci częściej klikali w reklamy. Na pewno warto też spróbować pomóc ludziom spać dłużej.

Przeciętny Amerykanin przesypia codziennie sześć i siedem dziesiątych godziny. Większość chciałaby więcej, ale przychodzi jedenasta wieczorem i zaczyna się *SportsCenter* albo wzywa nas YouTube, sen musi więc zaczekać. Jawbone, firma produkująca urządzenia pomiarowe i obsługująca setki tysięcy klientów, przeprowadza tysiące badań, by sprawdzić, w jaki sposób pomóc użytkownikom swoich produktów zrealizować marzenia: kłaść się spać trochę wcześniej.

I Jawbone odniosła wielki sukces, stawiając sobie niejako podwójny cel. Najpierw prosiła klientów, by zrealizowali niezbyt ambitny plan, wysyłając im wiadomość o następującej treści: „Wygląda na to, że nie wysypiałaś/wysypiałeś się za dobrze przez ostatnie trzy dni. Może spróbujesz położyć się dzisiaj o wpół do dwunastej? Wiemy, że zazwyczaj wstajesz o ósmej". Klient, jeśli tylko chce, klika opcję „Wchodzę w to".

Potem, kiedy przychodziła dziesiąta trzydzieści wieczorem, firma wysyłała klientom kolejną wiadomość: „Uzgodniliśmy, że spróbujesz się położyć o wpół do dwunastej. Dochodzi dziesiąta trzydzieści. A może spróbowałabyś/spróbowałbyś pójść spać już teraz?".

Okazało się, że dzięki takiej strategii klienci przesypiali średnio o dwadzieścia trzy minuty więcej niż zwykle. Firmie Jawbone nie udawało się położyć ludzi do łóżek o wpół do jedenastej, ale chodzili spać wcześniej.

Oczywiście każdy element tego schematu należało zoptymalizować doświadczalnie. Jeśli wyznaczymy pierwotny cel zbyt pochopnie –

poprosimy klientów, by kładli się spać o jedenastej wieczorem – to tylko bardzo nieliczni przystaną na taką propozycję. Podobnie niewiele osiągniemy, prosząc ludzi o to, by udawali się na spoczynek o północy.

Jawbone zastosowała testy A/B, chcąc znaleźć „senny ekwiwalent" słynnej strzałki wskazującej w prawo, użytej przez Google – tylko że zamiast zdobyć kilka kliknięć więcej dla reklamodawców wyszukiwarki, zyskała ponad dwadzieścia minut odpoczynku więcej dla przemęczonych Amerykanów.

Właściwie cała psychologia może korzystać z narzędzi z Doliny Krzemowej, żeby znacząco polepszyć wyniki badań. Niecierpliwie czekam na napisany przez jakiegoś psychologa pierwszy artykuł naukowy, w którym zamiast opisu dwóch doświadczeń przeprowadzonych z udziałem kilkorga studentów znajdą się wyniki tysiąca szybkich testów A/B.

Skończy się epoka, w której naukowcy tracili długie miesiące, żeby zatrudnić niewielką grupę młodzieży akademickiej do przeprowadzenia jednego testu. Teraz uczeni będą korzystać z danych cyfrowych, by sprawdzić kilkaset albo nawet kilka tysięcy pomysłów w ciągu zaledwie paru sekund. Zdobędziemy o wiele większą wiedzę w znacznie krótszym czasie.

Teksty jako dane również nauczą nas o wiele więcej. W jaki sposób rozprzestrzeniają się myśli i idee? Jak powstają nowe słowa? Jak znikają? Jak powstają dowcipy? Dlaczego niektóre słowa są śmieszne, a inne nie? Jak powstaje dialektyka? Jestem gotów się założyć, że za dwadzieścia lat będziemy dysponowali mądrymi odpowiedziami na te pytania.

Sądzę, że możemy też zastanowić się nad analizą zachowań dzieci w Internecie – oczywiście z poszanowaniem zasad anonimowości – która uzupełniałaby tradycyjne testy sprawdzające, jak nasi najmłodsi się uczą i ogólnie rozwijają. Jak radzą sobie z ortografią? Czy przejawiają oznaki dysleksji? Czy wykształcają w sobie dojrzałe zainteresowania intelektualne? Czy mają przyjaciół? Źródła odpowiedzi na te pytania kryją się w tysiącach uderzeń w klawiaturę, z której korzysta codziennie każde dziecko.

Poza tym jest jeszcze jeden niebanalny obszar, w którym możemy się spodziewać wielu rewelacji.

W piosence *Shattered* zespołu The Rolling Stones Mick Jagger śpiewa o tym, co sprawia, że Nowy Jork – zwany Wielkim Jabłkiem – jest takim magicznym miejscem. Śmiech. Radość. Samotność. Szczury. Pluskwy. Duma. Chciwość. Ludzie robiący sobie ubrania z papierowych toreb na zakupy. Ale Jagger najwięcej mówi o tym, co naprawdę nadaje temu miastu wyjątkowy charakter: „seks, seks, seks i seks".

Z Wielkim Jabłkiem jest tak jak z big data. Dzięki rewolucji cyfrowej znajdujemy odpowiedzi na pytania natury medycznej. Rozwiązujemy problemy dotyczące snu. Uczenia się. Psychologii. Języka. A także seksu, seksu, seksu i seksu.

Aktualnie zastanawiam się nad taką kwestią: ile właściwie istnieje wymiarów seksualności? Na ogół myślimy o ludziach po prostu jako o homo- albo heteroseksualnych, seksualność to jednak ewidentnie bardziej złożone zjawisko. Zarówno homo-, jak i heteroseksualiści mają przecież swoje preferencje – na przykład niektórzy mężczyźni wolą blondynki, a inni brunetki. Czy te upodobania manifestują się równie silnie, jak płciowe? Szukam też odpowiedzi na pytanie, skąd te nasze preferencje się biorą. Tak jak możemy ustalić, w jakim wieku kształtują się sympatie bejsbolowe czy poglądy polityczne, tak samo można już zbadać, w jakim wieku powstają inklinacje seksualne. Żeby się tego dowiedzieć, będziecie musieli kupić moją następną książkę, zatytułowaną roboczo *Wszyscy (dalej) kłamią*.

Pornografia, a szczególnie idące za nią dane, stała się rewolucyjnym czynnikiem w nauce o seksualności człowieka.

Trzeba było czasu, żeby nauki ścisłe zaczęły zmieniać nasze życie – zanim odkryliśmy penicylinę, zanim zaczęliśmy konstruować sztuczne satelity i komputery. I może minąć jeszcze sporo czasu, zanim big data poprowadzi nauki społeczne i behawioralne ku ważnym odkryciom, tłumaczącym, jak i dlaczego kochamy, uczymy się i żyjemy. Wierzę jednak, że takie odkrycia są blisko, i mam nadzieję, że po lekturze tej książki dostrzegacie chociaż ich zarysy. Więcej, mam nadzieję, że niektórzy z was się do nich przyczynią.

Aby ułożyć odpowiednie zakończenie, autor powinien zadać sobie pytanie, po co w ogóle napisał książkę. Jaki przyświecał mu cel?

Wydaje mi się, że napisałem *Wszyscy kłamią* pod wpływem jednej z najważniejszych lektur w moim życiu. Otóż niewiele ponad dziesięć lat temu ukazała się książka *Freakonomia*. Ten zdumiewający bestseller przedstawiał badania Stevena Levitta, wielokrotnie nagradzanego ekonomisty z University of Chicago, którego niejednokrotnie już wcześniej tu wspominałem. Levitt był „uczonym-chuliganem", który potrafił wykorzystywać dane analityczne, by odpowiedzieć na każde dziwaczne pytanie, jakie tylko przyszło mu do głowy: czy zapaśnicy sumo oszukują? Czy uczestnicy teleturniejów dyskryminują jedni drugich? Czy agenci nieruchomości przedstawiają klientom takie oferty, jakie sami wybraliby dla siebie?

W tamtym czasie świeżo ukończyłem studia filozoficzne i nie bardzo wiedziałem, co chcę robić w życiu. Ale kiedy przeczytałem *Freakonomię*, dotarło do mnie, że pójdę tą samą drogą, co Steven Levitt. Chciałem przekopywać się przez góry danych, żeby zrozumieć, jaki naprawdę jest świat. Postanowiłem, że będę robił to, co on, i napiszę doktorat z ekonomii.

Od tamtej pory minęło dwanaście lat i bardzo wiele się zmieniło. W kilku badaniach mojego idola znaleziono błędy w kodowaniu. Levitt wygłosił też parę niepoprawnych politycznie stwierdzeń na temat globalnego ocieplenia, a *Freakonomia* wypadła z łask w kręgach intelektualnych.

Myślę jednak, że jeżeli nie liczyć kilku pomyłek, to czas był życzliwy dla Levitta o tyle, o ile chodzi o pewne spostrzeżenie ogólniejszej natury, które zawarł w swej książce. Stwierdził bowiem, że połączenie ciekawości, kreatywności i analizy danych może nam bardzo pomóc zrozumieć świat. W danych analitycznych skrywały się opowieści, które trzeba było tylko z nich wydobyć, i to zostało dowiedzione już wielokrotnie.

Mam również nadzieję, że *Wszyscy kłamią* wywrze na czytelnikach podobne wrażenie, jakie *Freakonomia* zrobiła kiedyś na mnie. Liczę, że

moją książkę czyta w tej chwili jakiś młody człowiek, który sam nie za bardzo jeszcze wie, co począć ze swoim życiem. Jeśli znasz się trochę na statystyce, jesteś kreatywny i ciekawy świata, to zacznij pracować jako analityk danych.

Ośmielę się nawet powiedzieć, że można traktować moją książkę jako następczynię *Freakonomii*. Największą różnicą pomiędzy badaniami omawianymi tam i tutaj są ich ambicje. W latach dziewięćdziesiątych, gdy Levitt zdobył sławę, nie mieliśmy do dyspozycji tylu danych, a on chełpił się tym, że podejmuje różne dziwaczne problemy, jeśli tylko jest w stanie dotrzeć do danych na ich temat. Natomiast dziś, kiedy mamy mnóstwo materiałów źródłowych właściwie o wszystkim, należy raczej szukać odpowiedzi na ważne, ambitne pytania, aby móc sobie w końcu odpowiedzieć, co to znaczy być istotą ludzką.

Przyszłość analizy danych rysuje się jasno. Jestem prawie pewien, że następny Kinsey będzie analitykiem, podobnie jak kolejny Foucault, Freud, Marks, a nawet Salk.

Tak wyglądały zatem moje wysiłki zmierzające do napisania odpowiedniego zakończenia. Potem jednak zdałem sobie sprawę, że w wielkich zakończeniach chodzi o dużo więcej. O wiele, wiele więcej. Zrozumiałem, że dobre zakończenie powinno być ironiczne. Wzruszające. Poważne i wesołe jednocześnie. Głębokie, zabawne i smutne. Dobre zakończenie powinno w jednym czy dwóch zdaniach podsumowywać to, co powiedziało się wcześniej, jak również to, co pozostało jeszcze do powiedzenia. I powinno to zrobić w oryginalny, nowatorski sposób, przewrotnie. Dobra książka powinna się kończyć inteligentnym, śmiesznym i prowokacyjnym uderzeniem.

Więc może to dobry moment, żeby pomówić przez chwilę o moim pisaniu. Nie jestem autorem szczególnie elokwentnym. Niniejsza książka liczy zaledwie siedemdziesiąt pięć tysięcy słów, jest zatem dosyć krótka jak na tak obszerne zagadnienie, które porusza.

Ale to, czego mi brakuje w objętości, nadrabiam wytrwałością. Swój pierwszy felieton o seksie dla „New York Timesa", liczący dwa tysiące słów, pisałem przez pięć miesięcy w czterdziestu siedmiu wersjach.

Niektóre rozdziały *Wszyscy kłamią* miały sześćdziesiąt wersji. Potrafię poświęcić długie godziny na to, żeby znaleźć odpowiednie słowo, którego chcę użyć w jakimś zdaniu w przypisie.

Prawie przez cały ostatni rok żyłem jak pustelnik. Tylko ja i komputer. Mieszkam w najbardziej szpanerskiej dzielnicy Nowego Jorku, ale właściwie w ogóle nie wychodziłem z domu. Ta publikacja jest według mnie moim magnum opus, pomysłem mojego życia, byłem zatem gotów poświęcić wszystko, co trzeba, żeby ją napisać jak najlepiej. Chciałem móc obronić każde zawarte w niej słowo. Mój telefon pełen jest e-maili, na które zapomniałem odpowiedzieć, nieotwartych zaproszeń i wiadomości z aplikacji Bumble, które ignorowałem[23*].

Po trzynastu miesiącach ciężkiej pracy mogłem wreszcie posłać do wydawnictwa niemal kompletny tekst. Brakowało tylko jednej części: zakończenia.

Wytłumaczyłem mojej redaktorce, Denise, że zakończenie może mi zająć jeszcze kilka miesięcy. Najprawdopodobniej pół roku. Że moim zdaniem to najważniejsza część książki. I że dopiero zaczynam się orientować, jak dobre zakończenie powinno wyglądać. Nie muszę chyba dodawać, że Denise nie była zachwycona.

A potem pewnego dnia znajomy przysłał mi pocztą elektroniczną artykuł Jordana Ellenberga, matematyka z University of Wisconsin, który chciał się dowiedzieć, ilu ludzi czyta książki do końca, i wpadł na świetny pomysł, jak to sprawdzić, używając big data. Otóż Amazon podaje liczbę osób, które cytują różne fragmenty książek, a Ellenberg zrozumiał, że może porównać, ile przytoczeń pochodzi z początkowych, ile zaś z końcowych partii tekstu, co pozwoliłoby w przybliżeniu obliczyć, jak wielu z nas udaje się dobrnąć do końca lektury. Zgodnie z badaniami przeprowadzonymi tą metodą ponad dziewięćdziesiąt procent czytelników ukończyło powieść *Szczygieł* Donny Tartt[24*], lecz zaledwie siedem procent przedarło się przez magnum opus laureata Nagrody Nobla, ekonomisty Daniela Kahnemana *Pułapki myślenia. O myśleniu szybkim i wolnym*. Jeszcze mniej, bo trzy procent, przeczytało do końca szeroko dyskutowaną i wychwalaną książkę Thomasa

Piketty'ego *Kapitał w XXI wieku*[25*]. Wynikałoby z tego, że ludzie niechętnie kończą czytać dzieła napisane przez ekonomistów[4].

Jednym z wniosków płynących z mojej książki jest to, że musimy podążać za big data, dokądkolwiek nas prowadzi, i działać zgodnie ze wskazówkami płynącymi z analizy danych. Wolno mi żywić nadzieję, że większość z was będzie uważnie czytać każde moje słowo i postara się znaleźć konkretne zasady, na jakich ostatnie stronice łączą się z tym, co napisałem wcześniej. Ale bez względu na to, jak ciężko bym nie pracował nad wyszlifowaniem tekstu, przeważająca liczba czytelników przeczyta pierwszych pięćdziesiąt stron *Wszyscy kłamią*, zapamięta z nich kilka komentarzy i wróci do swojego codziennego życia jak gdyby nigdy nic.

Dlatego kończę w jedyny stosowny sposób: postępując tak, jak sugeruje analiza danych, czyli zgodnie z tym, co ludzie naprawdę robią, a nie z tym, co mówią. Idę zatem na piwo z przyjaciółmi i po prostu zostawiam całe to cholerne zakończenie w spokoju. Big data podpowiada mi, że na świecie jest za mało osób, które jeszcze w ogóle czytają książki.

PODZIĘKOWANIA

Ta książka to efekt zbiorowego wysiłku.

Przedstawione w niej koncepcje rodziły się, kiedy studiowałem na Harvardzie, i później, gdy pracowałem już jako analityk danych dla Google'a i pisałem felietony dla „New York Timesa".

Jedną z osób, które wywarły na mnie znaczny wpływ w tym okresie, był Hal Varian, mój współpracownik w Google'u. W moim najgłębszym przekonaniu Hal wyprzedza swój czas o jakieś dwadzieścia lat. Jego książka *Potęga informacji*[26*], napisana wspólnie z Carlem Shapiro, w zasadzie trafnie przewidziała przyszłość, a artykuł pod tytułem *Predicting the Present*, sygnowany także przez Hyunyoung Choi, w znacznej mierze zainicjował tę rewolucję big data w naukach społecznych, o której opowiadam w swojej książce. Hal jest też wspaniałym i sympatycznym mentorem, o czym zaświadczyć może wiele osób pracujących pod jego kierownictwem. Klasycznym posunięciem ze strony Hala jest wykonać większość roboty potrzebnej do napisania książki czy artykułu, nad którymi z nim pracujecie, a potem upierać się, żeby wasze nazwisko na okładce poprzedzało jego. Rzadko zdarzało mi się widzieć w życiu takie połączenie genialności i wielkoduszności.

Moje teksty i pomysły powstawały pod nadzorem Aarona Retika, redaktora wszystkich moich felietonów w „New York Timesie". Aaron jest omnibusem. Nie wiem, jak to możliwe, ale zna się na muzyce, historii, sporcie, polityce, socjologii, ekonomii i jeden Bóg wie, na czym jeszcze. To w znacznym stopniu jemu podpisane przeze mnie teksty w „Timesie" zawdzięczają to, co jest w nich dobrego. Do pozostałych zawodników w tej drużynie zaliczają się też Bill Marsh, którego grafiki nieustannie mnie powalają, Kevin McCarthy i Gita Daneshjoo. *Wszyscy*

kłamią zawiera fragmenty tych felietonów, wykorzystanych w książce za pozwoleniem redakcji „NYT".

Steve Pinker, który uprzejmie zgodził się napisać wstęp, od dawna jest moim bohaterem. To on ustawił wysoko poprzeczkę dla współczesnych książek z dziedziny nauk społecznych, prezentując fascynujący obraz podstaw wiedzy o naturze ludzkiej i przybliżając czytelnikom wyniki najciekawszych badań z wielu różnych dyscyplin. Do końca życia będę dążył do osiągnięcia tak wysokiego poziomu.

Rozprawę doktorską, będącą punktem wyjścia mojej książki, napisałem pod kierunkiem mądrych i cierpliwych konsultantów, do których należą Alberto Alesina, David Cutler, Ed Glaeser i Lawrence Katz.

Denise Oswald jest niesamowitą redaktorką. Jeżeli chcecie wiedzieć, jak dobrą, to porównajcie ostateczną wersję *Wszyscy kłamią* z pierwszą – chociaż właściwie nie będziecie mogli tego zrobić, bo tej kompromitującej pierwszej wersji nigdy już nikomu nie pokażę. Dziękuję również pozostałym członkom ekipy w wydawnictwie HarperCollins – Michaelowi Barrsowi, Lynn Grady, Lauren Janiec, Shelby Meizlik i Amber Oliver.

Mój agent, Eric Lupfer, od początku dostrzegał potencjał projektu książki o big data, sam sformułował propozycję dla wydawcy, a potem pomógł mi ją zrealizować.

Za wspaniałą robotę faktograficzną moje podziękowania niech zechce przyjąć Melvis Acosta.

Do pozostałych osób, dzięki którym wiele się nauczyłem w swoim życiu zawodowym i akademickim, należą: Susan Athey, Shlomo Benartzi, Jason Bordoff, Danielle Bowers, David Broockman, Bo Cowgill, Steven Delpome, John Donohue, Bill Gale, Claudia Goldin, Suzanne Greenberg, Shane Greenstein, Steve Grove, Mike Hoyt, David Laibson, A.J. Magnuson, Daba Maloney, Jeffrey Oldham, Peter Orszag, David Reiley, Jonathan Rosenberg, Michael Schwarz, Steve Scott, Rich Shavelson, Michael D. Smith, Lawrence Summers, Jon Vaver, Michael Wiggins i Quing Wu.

Dziękuję Timowi Requarthowi i NeuWrite za pomoc w pracy nad tekstem.

Z kolei za pomoc w interpretacji wyników różnych badań dziękuję Christopherowi Chabrisowi, Rajowi Chetty'emu, Mattowi Gentzkowowi, Solomonowi Messingowi i Jessemu Shapiro.

Spytałem kiedyś Emmę Pierson i Katię Sobolski, czy zechciałyby udzielić mi rad w kwestii jednego z rozdziałów. Z niezrozumiałych dla mnie powodów zaproponowały, że przeczytają całą książkę – a potem dzieliły się ze mną mądrymi uwagami na temat każdego akapitu.

Moja mama, Esther Davidowitz, czytała tekst w całości wielokrotnie i szalenie pomogła mi go ulepszyć. Nauczyła mnie też na własnym przykładzie, że powinienem iść za głosem swojej dociekliwości bez względu na to, dokąd mnie prowadzi. Kiedy poszedłem na rozmowę o pracę na uczelni, jeden z profesorów zaczął mnie grillować: „A co o tym, czym pan się zajmuje, myśli pańska matka?". Sugerował, że mama mogła czuć się skrępowana tym, że badam seks i inne tematy tabu. Ale ja zawsze wiedziałem, że mama jest dumna z tego, że chodzę wszędzie tam, dokąd mnie wiedzie ciekawość.

Fragmenty tej publikacji czytało wiele osób, które opatrywały ją następnie pomocnymi dla mnie komentarzami. Moje podziękowania za te konsultacje zechcą przyjąć Eduardo Acevedo, Coren Apicella, Sam Asher, David Cutler, Stephen Dubner, Christopher Glazek, Jessica Goldberg, Lauren Goldman, Amanda Gordon, Jacob Leshno, Alex Peysakhovich, Noah Propp, Ramon Roullard, Greg Sobolski, Evan Soltas, Noah Stephens-Davidowitz, Lauren Stephens-Davidowitz i Jean Yang. Jean była właściwie moją najlepszą przyjaciółką, kiedy pisałem *Wszyscy kłamią*, dziękuję jej więc również za tę przyjaźń.

Natomiast za pomoc w zbieraniu danych dziękuję Brettowi Goldenbergowi, Jamesowi Rogersowi i Mike'owi Williamsowi z Mind-Geek, a także Robowi McQuownowi i Samowi Millerowi z Baseball Prospectus.

Za wsparcie finansowe wyrazy wdzięczności zechce przyjąć Alfred Sloan Foundation.

Na pewnym etapie pracy stanąłem w miejscu, pogubiłem się i byłem bliski rezygnacji z projektu. Wyjechałem wtedy na wieś z moim tatą, Mitchellem Stephensem, i dzięki niemu w ciągu tygodnia stanąłem znów na nogi. Ojciec zabierał mnie na spacery, na których rozmawialiśmy o miłości, śmierci, sukcesach, szczęściu i pisaniu – a kiedy wracaliśmy do domu, kazał mi siadać do tekstu, żebyśmy mogli przyjrzeć się uważnie każdemu jego zdaniu. Nie ukończyłbym tej książki bez niego.

Za wszystkie znajdujące się jeszcze w tekście błędy odpowiedzialność ponoszę, rzecz jasna, wyłącznie ja.

PRZYPISY

WSTĘP

[1] Katie Fretland, *Gallup: Race Not Important to Voters*, The Swamp, „Chicago Tribune", VI 2008.

[2] Alexandre Mas i Enrico Moretti, *Racial Bias in the 2008 Presidential Election*, „American Economic Review" 2009, nr 2.

[3] 12 XI 2009 roku Lou Dobbs stwierdził w swoim programie telewizyjnym, że żyjemy w społeczeństwie „postkontestacyjnym i postrasowym". Z kolei 27 I 2010 roku w programie Chrisa Matthewsa usłyszeliśmy oświadczenie autora, że wszystko wskazuje na to, iż Barack Obama „jest prezydentem postrasowym". Więcej podobnych przykładów podają Michael C. Dawson i Lawrence D. Bobo, *One Year Later and the Myth of a Post-Racial Society*, „Du Bois Review: Social Science Research on Race" 2009, nr 2.

[4] Szczegóły wszystkich tych wyliczeń można znaleźć na mojej stronie sethsd.com, w zakładce pod nazwą „Sex Data". Dane z General Social Survey znajdują się pod adresem http://gss.norc.org/.

[5] Dane przekazane autorowi.

[6] Autorska analiza danych z Google Trends. Przejrzałem też dane wszystkich członków Stormfrontu i napisałem o tym w artykule pod tytułem *Data of Hate*, „New York Times", 13 VII 2014, SR4. Informacje, o których mowa, można pobrać z sethsd.com spod zakładki „Stormfront".

[7] Autorska analiza danych z Google Trends. Stany, dla których to twierdzenie jest prawdziwe, to między innymi Kentucky, Luizjana, Arizona i Karolina Północna.

[8] Artykuł ten został ostatecznie opublikowany w „Journal of Public Economics" 2014, nr 118, pod tytułem *The Cost od Racial Animus on a Black Candidate: Evidence Using Google Search Data*. Można w nim znaleźć więcej szczegółów z tych badań. Dodatkowe informacje podaję też na stronie sethsd.com w zakładce „Racism".

[9] „Najsilniejszy korelat dla poparcia Trumpa, jaki udało mi się znaleźć, to wyszukiwania w Google'u słowa na «cz». Donosili też o tym inni badacze" (28 II 2016, tweet). Zob. też Nate Cohn, *Donald Trump's Strongest Supporters: A New Kind of Democrat*, „New York Times", 31 XII 2015, A3.

Należy zauważyć, że ponieważ jednostką miary jest procent zapytań w Google'u, nie jest on arbitralnie wyższy w miejscach bardziej zaludnionych czy miejscach, skąd pochodzi najwięcej wyszukiwań. Warto również wskazać, że niektóre różnice pomiędzy tą mapą a mapą poparcia dla Trumpa da się łatwo wyjaśnić. Trump stracił poparcie w Teksasie i Arkansas, ponieważ z tych stanów wywodzą się jego dwaj oponenci, Ted Cruz i Mike Huckabee.

Dane pochodzą z Civis Analytics z grudnia 2015. Faktyczne dane dotyczące głosowania są w tym przypadku mniej użyteczne, ponieważ wpłynął na nie termin prawyborów oraz metoda głosowania. Mapy przedrukowano za zgodą „New York Timesa".

[10] *Bringing Big Data to the Enterprise*, IBM, https://www-01.obm.com/software/data/bigdata/what-is-big-data.html.

[11] Nassim M. Taleb, *Beware the Big Errors of „Big Data"*, „Wired", 8 II 2013, http://www.wired.com/2013/02/big-data-means-big-errors-people.

[12] Zbadałem, w jaki sposób rasizm internetowy zmieniał się w regionach kraju mocniej i słabiej dotkniętych kryzysem ekonomicznym. Przyjrzałem się zarówno liczbie wpisów typu

„czarnuch(y)", jak i liczbie nowych członków Stormfrontu. Dane te można pobrać ze strony sethsd.com w zakładkach „Racial Animus" i „Stormfront".

[13] Seth Stephens-Davidowitz, *Fifty States of Anxiety*, „New York Times", 7 VIII 2016, SR2. Proszę zauważyć, że chociaż wyszukiwania w Google'u rzeczywiście stanowią obszerniejszy materiał badawczy, to prawidłowość tę potwierdzają dane z sondaży i ankiet. Zob. na przykład William C. Reeves et al., *Mental Illness Surveillance Among Adults in the United States*, „Morbidity and Mortality Weekly Report Supplement" 2011, nr 3.

[14] Omawiam to zagadnienie w artykule *Why Are You Laughing?*, „New York Times", 15 V 2016, SR9. Dane można pobrać z sethsd.com, z zakładki pod nazwą „Jokes".

[15] Zjawisko to przedstawiam w artykule *What Do Pregnant Women Want?*, „New York Times", 17 V 2014, SR6.

[16] Autorska analiza danych z portalu PornHub.

[17] Piszę na ten temat w artykule *Searching for Sex*, „New York Times", 25 I 2015, SR1.

[18] Seth Stephens-Davidowitz, *What Do Pregnant Women Want?*, *op.cit.*

[19] Rozmawiałem z Jerrym Friedmanem przez telefon 27 X 2015.

[20] Hal R. Varian, *Big Data: New Tricks for Econometrics*, „Journal of Economic Perspectives" 2014, nr 2.

ROZDZIAŁ PIERWSZY: NASZA OMYLNA INTUICJA

[1] Mówię o tym aspekcie analizy danych, na którym się znam, czyli o analizie tłumaczącej i prognozującej zachowanie człowieka, a nie o sztucznej inteligencji, dzięki której maszyna mogłaby na przykład prowadzić samochód. Taką technologię, mimo że korzysta ona

z narzędzi wzorowanych na działaniu ludzkiego mózgu, oczywiście trudniej jest zrozumieć.

[2] John Paparizos, Ryan W. White, Eric Horvitz, *Screening for Pancreatic Adenocarcinoma Using Signals from Web Search Logs: Feasibility Study and Results*, „Journal of Oncology Practice" 2016.

[3] Omawiam te badania w swoim artykule *Dr. Google Will See You Now*, „New York Times", 11 VIII 2013, SR12.

[4] Lars Backstrom, John Kleinberg, *Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook*, w: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2014).

[5] Daniel Kahneman, *Pułapki myślenia. O myśleniu szybkim i wolnym*, przeł. Piotr Szymczak, Poznań 2012.

[6] Między 1979 i 2010 rokiem w wyniku tornad zmarło statystycznie 55,81 Amerykanów, a na astmę 4216,53. Zob. *Annual U.S. Killer Tornado Statistics*, National Weather Service, http://www,spc.noaa.gov./climo/torn/fatalmap.php i *Trends in Asthma Morbidity and Mortality*, American Lung Association, Epidemiology and Statistics Unit.

[7] Moje ulubione filmy wideo z Ewingiem to *Patrick Ewings's Top 10 Career Plays*, film na YouTube, 18 IX 2015, http://www.youtube.com/watch?v=Y29gMuYymv8, i *Patrick Ewing Knicks Tribute*, film na YouTube, post z 12 V 2006, http://www.youtube.com/watch?v=8T2l5Emzu-I.

[8] S.L. Price, *Whatever Happened to the White Athlete?*, „Sports Illustrated", 8 XII 1997.

[9] Ankieta przeprowadzona w ramach Google Consumer Survey 22 X 2013. Moje pytanie brzmiało: „Jak się wam wydaje, gdzie urodziła się większość graczy z NBA?". Ankietowani mieli do wyboru dwie odpowiedzi: „w dzielnicach/regionach biedy"

i „w dzielnicach/regionach zamieszkanych przez klasę średnią"; 59,7 procent respondentów wybrało pierwszą odpowiedź.

[10] Roland G. Fryer Jr, Steven D. Levitt, *The Causes and Consequences of Distinctively Black Names*, „Quarterly Journal of Economics" 2004, nr 3.

[11] Center for Disease Control and Prevention, „Health, United States, 2009", tablica 9, *Nonmarital Childbearing, by Detailed Race and Hispanic Origin of Mother, and Maternal Age: United States, Selected Years 1970–2006*.

[12] *Not Just a Typical Jock: Miami Heat Forward Chris Bosh's Interests Go Well Beyond Basketball*, palmbeachpost.com, 15 II 2011, http://www.palmbeachpost.com/news/sports/basketball/not-just-a-typical-jock-miami-heat-forward-chris-b/nLp7Z/; Dave Walker, *Chris Paul's Family to Compete on Family Feud*, nola.com, 31 X 2011, http://www.nola.com/tv/index.ssf/2011/10/chris-pauls-family-to-compete.html.

[13] *Why Are We Getting Taller as a Species?*, „Scientific American", http://scientificamerican.com/article/why-are-we-getting-taller/. Co ciekawe, Amerykanie już chyba przestali dziś rosnąć (Amanda Onion, *Why Have Americans Stopped Growing Taller?*, ABC News, 3 VII 2016, http://abcnews.go.com/Technology/story?id=98438&page=1). Moim zdaniem jednym z powodów, dla których mamy dziś w NBA tak wielu graczy urodzonych za granicą, jest właśnie to, że mężczyźni w innych krajach doganiają Amerykanów, jeżeli chodzi o wzrost. Liczba urodzonych w Stanach Zjednoczonych zawodników NBA, mierzących siedem stóp albo więcej, zwiększyła się szesnastokrotnie pomiędzy 1946 i 1980 rokiem, kiedy wzrost mieszkańców USA się zwiększał, od tamtej pory przestała jednak piąć się w górę wraz z Amerykanami. Zwiększyła się natomiast liczba bardzo wysokich graczy z innych państw. Najwięcej z nich, jak stwierdziłem, pochodzi z Turcji, Hiszpanii i Grecji, gdzie w ostatnich latach polepszyła się opieka zdrowotna nad dziećmi i odnotowano wyższy przeciętny wzrost wśród dorosłych mężczyzn.

[14] Carmen R. Isasi et al., *Association of Childhood Economic Hardship with Adult Height and Adult Adiposity among Hispanics/Latinos, The HCHS/SOL Socio-Cultural Ancillary Study*, „PloS One" 2016, nr 2; Jane E. Miller,

Sanders Korenman, *Poverty and Children's Nutrition Status in the United States*, „American Journal of Epidemiology" 1994, nr 3; Harry J. Holzer, Diane Whitmore Schanzenbach, Greg J. Duncan, Jens Ludwig, *The Economic Costs of Childhood Poverty in the United States*, „Journal of Children and Poverty" 2008, nr 1.

[15] Cheryl D. Fryar, Qiuping Gu, Cynthia L. Ogden, *Anthropometric Reference Data for Children and Adults: United States, 2007–2010*, „Vital and Health Statistics Series" 2012, nr 252.

[16] Pablo S. Torre, *Larger Than Real Life*, „Sports Illustrated", 4 VII 2011.

[17] Tim Kautz, James J. Heckman, Ron Diris, Bas Ter Weel, Lex Borghans, *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*, National Bureau of Economic Research Working Paper 20749, 2014.

[18] Desmond Conner, *For Wrenn, Sky's the Limit*, Hartford Cournat, 21 X 1999.

[19] Historię Douga Wrenna opisuje Percy Allen, *Former Washington and O'Dea Star Doug Wrenn Finds Tough Times*, „Seattle Times", 29 III 2009.

[20] *Ibidem*.

[21] Melissa Isaacson, *Portrait of a Legend*, espn.com, 9 IX 2009, http://www.espn.com/chicago/columns/story?id=4457017&columnist=isaacson_melissa. Dobrą biografię Jordana napisał Roland Lazenby (wyd. polskie: Roland Lazenby, *Michael Jordan. Życie*, przeł. Michał Rutkowski, Kraków 2014).

[22] Barry Jacobs, *High Flying Michael Jordan Has North Carolina Cruising Toward Another NCAA Title*, „People", 19 III 1984.

[23] Isaacson, *op.cit*.

[24] *Michael Jordan's Basketball Hall of Fame Enshrinment Speech*, YouTube, 21 II 2012, http://www.youtube.com/watch?v=XLzBMGXfK4c. Najciekawszym aspektem przemówienia Jordana nie jest to, że tak serdecznie traktuje swoich rodziców, tylko że ciągle odczuwa potrzebę

wspominania porażek z początków swojej kariery. Być może trwająca całe życie obsesja na tym punkcie jest potrzebna, by zostać największym zawodnikiem w historii koszykówki.

[25] *I'm LeBron James from Akron, Ohio*, YouTube, 20 VI 2013, http://www.youtube.com/watch?v=XceMbPVAggk.

ROZDZIAŁ DRUGI: CZY FREUD MIAŁ RACJĘ?

[1] Za falliczne w kształcie uznałem te owoce i warzywa, które są wyraźnie dłuższe niż szersze i raczej owalne, w tym ogórki, kukurydzę w kolbach, marchewki, oberżyny, podłużne dynie i banany.

[2] Zbiór ten można pobrać pod adresem https://www.microsoft.com/en-us/download/details.aspx?id=52418. Badacze prosili użytkowników Amazon Mechanical Turk o opis zdjęć, a potem analizowali poszczególne uderzenia w klawisze i odnotowywali każdą korektę ortograficzną. Więcej szczegółów można znaleźć w: Yukino Baba, Hisami Suzuki, *How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs*, w: *Proceedings of the Fiftieth Annual Meeting of the Association for Computational Linguistics*, 2012.

[3] Pełne dane – uwaga: dość drastyczne – wyglądają następująco:

Chcę uprawiać seks z moim/moją...

	Liczba miesięcznych wpisów w Google dokładnie o tej treści
mamą	720
synem	590
siostrą	590
kuzynem/kuzynką	480

tatą	480
chłopakiem	480
bratem	320
córką	260
przyjacielem/przyjaciółką	170
dziewczyną	140

[4] Na przykład „porno” jest jednym ze słów najczęściej pojawiających się w internetowych wyszukiwaniach różnych niezwykle popularnych filmów animowanych, co widać w poniższej tabeli.

Animacja i porno
(najpopularniejsze wyszukiwania filmów animowanych)

family guy porno	oglądaj the simpsons	**futurama porno**	scooby doo gry
family guy odcinki	**the simpsons porno**	futurama leela	scooby doo fabuła
family guy za darmo	the simpsons online	futurama odcinki	**scooby doo porno**
oglądaj family guy	the simpsons fabuła	futurama online	scooby doo velma

[5] Oto dziewięć kobiecych zawodów poszukiwanych najczęściej (według moich wyliczeń) na stronach pornograficznych przez mężczyzn z trzech grup wiekowych:

Zawody kobiece w męskich wyszukiwaniach pornograficznych, trzy grupy wiekowe

	18–24	**25–64**	**65+**
1.	Opiekunka do dzieci	Opiekunka do dzieci	Opiekunka do dzieci
2.	Nauczycielka	Instruktorka jogi	Czirliderka
3.	Instruktorka jogi	Nauczycielka	Lekarka

4.	Czirliderka	Czirliderka	Nauczycielka
5.	Lekarka	Agentka nieruchomości	Agentka nieruchomości
6.	Prostytutka	Lekarka	Pielęgniarka
7.	Agentka nieruchomości	Prostytutka	Instruktorka jogi
8.	Pielęgniarka	Sekretarka	Sekretarka
9.	Sekretarka	Pielęgniarka	Prostytutka

ROZDZIAŁ TRZECI: DANE PRZEOBRAŻONE

[1] Matthew Leising, *HFT Treasury Trading Hurts Market When News Is Released*, „Bloomberg Markets", 16 XII 2014; Nathaniel Popper, *The Robots Are Coming for Wall Street*, „New York Times Magazine", 28 II 2016, MM56; Richard Finger, *High Frequency Trading: Is It a Dark Force Against Ordinary Human Traders and Investors?*, „Forbes", 30 IX 2013, http://www.forbes.com/sites/richardfinger/2013/09/30/high-frequency-trading-is-it-a-dark-force-against-ordinary-human-traders-and-investors/#50875fc751a6.

[2] Rozmawiałem z Alanem Kruegerem przez telefon 8 V 2015.

[3] Pierwszą pracę na ten temat opublikowali Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski i Larry Brilliant, *Detecting Influenza Epidemics Using Search Engine Query Data*, „Nature" 2009, nr 7232. Wady pierwszego algorytmu opisali David Lazer, Ryan Kennedy, Gary King i Alessandro Vespigniani, *The Parable of Google Flu: Traps in Big Data Analysis*, „Science" 2014, nr 6176. Ulepszony algorytm przedstawili Shihao Yang, Mauricio Santillana i S.C. Kou, *Accurate Estimation of Influenza Epidemics Using Google Search Data Via ARGO*, „Proceedings of the National Academy of Sciences" 2015, nr 47.

[4] Seth Stephens-Davidowitz, Hal Varian, *A Hands-on Guide to Google Data*, mimeo, 2015.

[5] Sergey Brin i Larry Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Seventh International World-Wide Web Conference, 14–18 IV 1998, Brisbane, Australia.

[6] John Battelle, *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, New York 2005.

[7] Dobre omówienie tego zagadnienia można znaleźć w książce Stevena Levy'ego, *In the Plex: How Google Thinks, Works and Shapes Our Lives*, New York 2011.

[8] Cytat ten można znaleźć także w: Joe Drape, *Ahmed Zayat's Journey: Bankruptcy and Big Bets*, „New York Times", 5 VI 2015, A1, gdzie jednak został mylnie przypisany Sederowi, słów tych użył bowiem w rzeczywistości jeden z członków jego zespołu.

[9] Rozmawiałem z Jeffem Sederem i Patty Murray w Ocala na Florydzie w dniach 12–14 VI 2015.

[10] Tak przyczyny porażek koni wyścigowych ocenia Jeff Seder, opierając się na swoich wieloletnich doświadczeniach zawodowych.

[11] Supplemental Tables of Equine Injury Database Statistics for Thoroughbreds, http://jockeyclub.com/pdfs/eid_7_year_tables.pdf.

[12] *Postmortem Examination Program*, California Animal Health and Food Laboratory System, 2013.

[13] Avalyn Hunter, *A Case for Full Siblings*, „Bloodhorse", 18 IV 2014, http://www.bloodhorse.com/horse-racing/articles/115014/a-case-for-full-siblings.

[14] Melody Chiu, *E.J. Johnson Loses 50 Lbs. since Undergoing Gastric Sleeve Surgery*, „People", 1 X 2014.

[15] Eli Saslow, *Lost Stories of LeBron, Part 1*, espn.com, 17 X 2013, http://www.espn.com/nba/story/_/id/9825052/how-lebron-james-life-changes-fourth-grade-espn-magazine.

[16] Zob. Sherry Ross, *16 Million Dollar Baby*, „New York Daily News", 12 III 2006, i Jay Privman, *The Green Monkey, Who Sold for $16M, Retired*, espn.com, 12 II 2008, http://www.espn.com/sports/horse/news/story?id=3242341. Nagranie wideo z tej aukcji można zobaczyć pod adresem *$16 Million Horse* na YouTube, 1 XI 2008, http://www.youtube.com/watch?v=EyggMC85Zsg.

[17] Sharad Goel, Jake M. Hoffman, Sébastien Lahaie, David M. Pennock, Duncan J. Watts, *Predicting Consumer Behavior with Web Search*, „Proceedings of the National Academy of Sciences" 2010, nr 41.

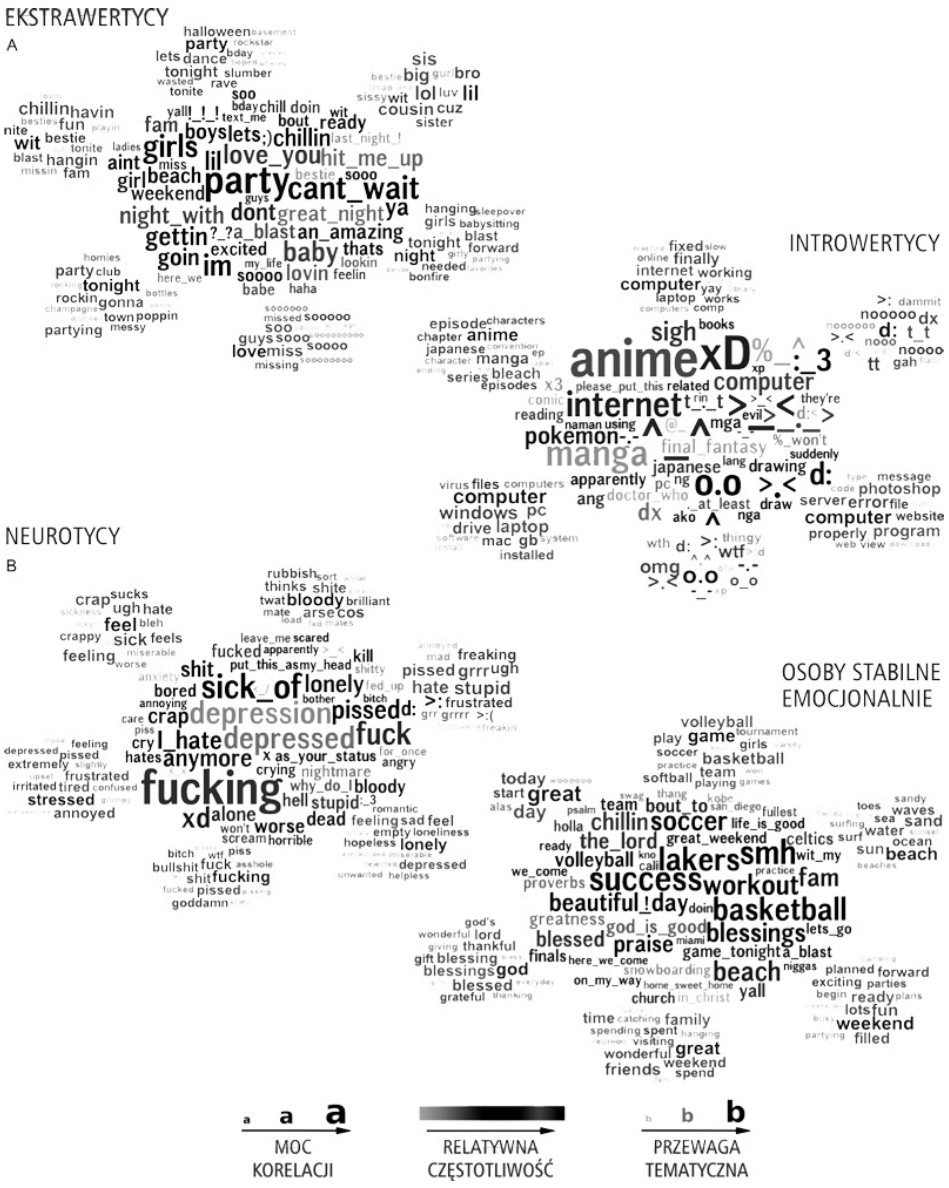
[18] Constance L. Hays, *What Wal-Mart Knows About Customers' Habits*, „New York Times", 14 XI 2004.

[19] Rozmawiałem z Orleyem Ashenfelterem przez telefon 27 X 2016.

[20] Daniel A. McFarland, Dan Jurafsky, Craig Rawlings, *Making the Connection: Social Bonding in Courtship Situations*, „American Journal of Sociology" 2013, nr 6.

[21] Jonathan Greenberg, *What I Learned From My Wise Uncle Leonard Cohen*, „Huffington Post", 11 XI 2016.

[22] H. Andrew Schwartz et al., *Personality, Gender and Age in the Language of Social Media: The Open-Vocabulary Approach*, „PloS One" 2013, nr 9. Artykuł omawia także różnice w słownictwie badanych, których pogrupowano według wyników ich testów osobowościowych. Oto rezultaty analizy autorów:



[23] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, Peter Sheridan Dodds, *The Emotional Arcs of Stories Are Dominated by Six Basic Shapes*, „EPJ Data Science" 2016, nr 1.

[24] Jonah Berger, Katherine L. Milkman, *What Makes Online Content Viral?*, „Journal of Marketing Research" 2012, nr 2.

[25] Badania te zostały szczegółowo omówione w: Matthew Gentzkow, Jesse M. Shapiro, *What Drives Media Slant? Evidence from U.S. Daily Newspapers*, „Econometrica" 2010, nr 1. Gentzkow i Shapiro byli zaledwie doktorantami, gdy zaczynali ten projekt, dzisiaj są gwiazdami

ekonomii. Gentzkow, profesor na Stanford University, zdobył w 2014 roku John Bates Clark Medal, przyznawany najlepszym ekonomistom, którzy nie ukończyli czterdziestego roku życia. Shapiro, profesor na Brown University, jest redaktorem prestiżowego pisma „Journal of Political Economy". Ich wspólny artykuł o politycznej stronniczości mediów do dziś pozostaje najczęściej cytowanym tekstem w wypadku obu uczonych.

[26] To, że Murdoch jest właścicielem konserwatywnego „New York Post", można wyjaśnić wielkością Nowego Jorku: w tak dużym mieście znajdzie się miejsce dla gazet reprezentujących najróżniejsze poglądy. Wydaje się jednak dosyć oczywiste, że „Post" regularnie przynosi straty. Zob. na przykład Joe Pompeo, *How Much Does the „New York Post" Actually Lose*, „Politico", 30 VIII 2013, http://www.politico.com./media/story/2013/08/how-much-does-the-new-york-post-actually-lose-001176.

[27] Rozmawiałem z Mattem Gentzkowem i Jessem Shapiro 16 VIII 2015 w Royal Sonesta w Bostonie.

[28] Kate Rakelly, Sarah Sachs, Brian Yin, Alexei A. Efros, *A Century of Portraits: A Visual Historical Record of American High School Yearbooks*, artykuł zaprezentowany na International Conference on Computer Vision, 2015. Zdjęcia przedrukowuję za pozwoleniem autorów.

[29] Zob. na przykład Christina Kotchemidova, *Why We Say „Cheese": Producing the Smile in Snapshot Photography*, „Critical Studies in Media Communication" 2005, nr 1.

[30] J. Vernon Henderson, Adam Storeygard, David N. Weil, *Measuring Economic Growth from Outer Space*, „American Economic Review" 2012, nr 2.

[31] Kathleen Caulderwood, *Nigerian GPD Jumps 89% As Economists Add in Telecoms, Nollywood*, „IBTimes", 4 VII 2014, http://www.ibtimes.com/nigerian-gdp-jumps-89-economists-add-telecoms-nollywood-1568219.

[32] Rozmawiałem z Joem Reisingerem przez telefon 10 VI 2015.

[33] Leena Rao, *Space X and Tesla Backer Just Invested $50 Million in This Startup*, „Fortune", 24 IX 2015.

ROZDZIAŁ CZWARTY: CYFROWE SERUM PRAWDY

[1] Hugh J. Parry, Helen M. Crossley, *Validity of Responses to Survey Questions*, „Public Opinion Quarterly" 1950, nr 1.

[2] Frauke Kreuter, Stanley Presser, Roger Tourangeau, *Social Desirability Bias in CATI, IVR and Web Surveys*, „Public Opinion Quarterly" 2008, nr 5.

[3] Artykuł dowodzący, że prognozy poparcia dla Donalda Trumpa mogło wypaczyć kłamstwo, napisał Thomas B. Edsall, *How Many People Support Trump But Don't Want to Admit It?*, „New York Times", 15 V 2016, SR2. Aby zapoznać się z opinią, że kłamstwo nie miało w tej kwestii większego znaczenia, zob. Andrew Gelman, *Explanations for That Shocking 2% Shift*, „Statistical Modeling, Causal Inference and Social Science", 9 XI 2016, http://anfrewgelman.com/2016/11/09/explanations-shocking-2-shift/.

[4] Rozmawiałem z Rogerem Tourangeau przez telefon 5 V 2015.

[5] Zagadnienie to omawia Adam Grant w książce *Originals: How Non-Conformists Move the World* (New York 2016). Pierwszym źródłem, jeśli chodzi o to zagadnienie, był dla mnie artykuł Davida Dunninga, Chipa Heatha i Jerry'ego M. Sulsa, *Flawed Self-Assessment: Implications for Health, Education and the Workplace*, „Psychological Science in the Public Interest" 2004, nr 5.

[6] Anya Kamenetz, *Mischievous Responders Confound Research on Teens*, „nprED", 22 V 2014, http://www.npr.org/sections/ed/2014/05/22/313166161/mischievous-responders-confound-research-on-teens. Artykuł ten omawia

wcześniejsze badania, które przedstawił Joseph P. Robinson-Cimpian, *Inaccurate Estimation of Disparities Due to Mischievous Responders*, „Educational Researcher" 2014, nr 4.

[7] https://www.google.com/trends/explore?date=all&geo=US&q+porn,weather.

[8] Amanda Hess, *How Many Women Are Not Admitting to Pew That They Watch Porn?*, 11 X 2013, „Slate", http://www.slate.com/blogs/xx_factor/2013/10/11/pew-_online-viewing-study-percentage-of-women-who-watch-online-porn-is-growing.html.

[9] Nicholas Diakopoulus, *Sex, Violence and Autocomplete Algorithms*, „Slate", 2 VIII 2013, http://www.slate.com/articles/technology/future_tense/2013/08/words-banned-from-bing-and-google-s-autocomplete-algorithms.html.

[10] Na podstawie różnych sformułowań szacuję, że w amerykańskim Google'u co miesiąc pojawia się mniej więcej 1730 wyszukiwań, z których wyraźnie wynika, że ich autorzy żałują, że mają dzieci. Ubolewanie z powodu braku potomstwa wyraża natomiast jedynie około 50 osób. W Stanach Zjednoczonych mieszka blisko 16 milionów ludzi powyżej czterdziestu pięciu lat, którzy nie mają dzieci, ma je natomiast około 152 milionów Amerykanów. Oznacza to, biorąc pod uwagę dorosłą populację, że osoby będące rodzicami żalą się w Internecie trzy i sześć dziesiątych razy częściej niż bezdzietne. Oczywiście, choć była już o tym mowa w tekście głównym, warto jednak powtórzyć, że takie wyznania w Google'u czyni tylko niewielka liczba internautów, zapewne rozgoryczonych swoimi dziećmi na tyle, aby zapomnieć, że w tych sprawach wyszukiwarka nie może im w niczym pomóc.

[11] Wyliczenia te pochodzą z pracy Nate'a Silvera, *How Opinion on Same-Sex Marriage Is Changing, and What It Means*, FiveThirtyEight, 26 III 2013, http://fivethirtyeight.blogs.nytimes.com/2013/03/26/how-opinion-on-same-sex-marriage-is-changing-and-what-it-means/?_r=0.

[12] Jak już wspominałem, Google Trends nie grupuje wyszukiwań podług płci ich autorów. Owszem, Google AdWords odnotowuje płeć internautów przeglądających określone strony, ale dane te są o wiele mniej precyzyjne. Chcąc sprawdzić liczbę wpisów według płci, korzystam najpierw z danych z wyszukiwarki, aby orientacyjnie określić odsetek wyszukiwań gejowskich w każdym stanie, a następnie precyzuję wyniki, zestawiając je z danymi „płciowymi" z Google AdWords. Inny sposób na zdobycie danych rozgraniczonych według kryterium płci to odwołanie się do informacji z PornHubu. Portal ten jednak dostarcza najprawdopodobniej mocno wyselekcjonowane dane, bo wielu gejów pewnie raczej woli korzystać z portali poświęconych wyłącznie pornografii homoseksualnej. PornHub sugeruje, że mężczyźni korzystają z gejowskiego porno rzadziej, niż wynikałoby to z analizy Google'a. Potwierdza jednak, że nie istnieje silny związek między tolerancją wobec homoseksualizmu i korzystaniem z gejowskich treści pornograficznych. Wszystkie te dane i dalsze komentarze znajdują się na mojej stronie sethsd.com w zakładce „Sex".

[13] Wyliczenia autora na podstawie danych facebookowych: 8 lutego 2017 ok. 300 licealistów płci męskiej w San Francisco-Oakland_San Jose miało na profilach zamieszczoną informację, że interesują ich mężczyźni, a ok. 7800 napisało, że interesują ich kobiety.

[14] *„We Don't Have Any Gays in Iran", Iranian President Tells Ivy League Audience*, „Daily Mail.com.", 25 IX 2007, http://www.dailymail.co.uk/news/article-483746/We-dont-gays-Iran-Iranian-president-tells-Ivy-League-audience.html.

[15] Brett Logiurato, *Sochi Mayor Claims There Are No Gay People in the City*, „Sports Illustrated", January 27, 2014.

[16] Zgodnie z danymi z Google AdWords w Soczi co roku pojawiają się dziesiątki wyszukiwań *гей порно* (gejowskiej pornografii). Procentowo rzecz biorąc, jest ich mniej więcej tyle samo, co w Stanach Zjednoczonych. Google AdWords nie zawiera danych z Iranu, podobnie jak PornHub, dane irańskie przejrzała jednak wyszukiwarka PornMD, która podaje, że pięć z dziesięciu najpopularniejszych wpisów w Iranie

dotyczyło pornografii gejowskiej. Znalazły się tam takie hasła jak „miłość tatusia" i „biznesmen z hotelu", a opisał je Joseph Patrick McCormick w artykule *Survey Reveals Searches for Gay Porn Are Top in Countries Banning Homosexuality*, „Pink News", http://www.pinknews.co.uk/2013/03/13/survey-reveals-searches-for-gay-porn-are-top-in-countries-banning-homosexuality/. Według Google Trends jakieś dwa procent wyszukiwań pornograficznych w Iranie dotyczy pornografii gejowskiej, czyli mniej niż w Stanach Zjednoczonych, oznacza to jednak, że zainteresowanie takimi treściami jest duże.

[17] Seth Stephens-Davidowitz, *Searching for Sex*. Dane do tej części książki można znaleźć na mojej stronie sethsd.com w zakładce „Sex".

[18] *Current Contraceptive Status Among Women Aged 15–44: United States, 2011–2013*, Centers for Disease Control and Prevention, http://www.cdc.gov/nchs/data/databriefs/db173_table.pdf#1.

[19] David Spiegelhalter, *Sex: What Are the Chances?*, BBC News, 15 III 2012, http://www.bbc.com/future/story/20120313-sex-in-the-city-or-elsewhere.

[20] W Stanach Zjednoczonych mamy około sześciu milionów sześciuset tysięcy ciąż rocznie i 62 miliony kobiet w wieku od piętnastu do czterdziestu czterech lat.

[21] Jak wspominałem, nie znam płci autorów wpisów w Google'u, zakładam więc, że przeważająca większość pytań dotyczących sposobów dostarczania oralnej przyjemności kobiecie pochodzi od mężczyzn, a przeważającą większość pytań dotyczących sposobów dawania oralnej przyjemności mężczyźnie zadają kobiety. Przyjmuję to założenie zarówno dlatego, że ogromna większość społeczeństwa jest heteroseksualna, jak i dlatego, że ludzie generalnie chyba rzadziej chcą wiedzieć, jak sprawić seksualną rozkosz osobie tej samej płci.

[22] Autorska analiza danych z Google AdWords.

[23] Evan Soltas, Seth Stephens-Davidowitz, *The Rise of Hate Search*, „New York Times", 13 XII 2015, SR1. Dane na ten temat i więcej szczegółów na mojej stronie internetowej sethsd.com w zakładce „Islamophobia".

[24] Autorska analiza danych z Google Trends.

[25] Autorska analiza danych z Google Trends.

[26] Ashwin Rode i Anand J. Shukla, *Prejudicial Attitudes and Labor Market Outcomes*, mimeo, 2013.

[27] Seth Stephens-Davidowitz, *Google, Tell Me. Is My Son a Genius?*, „New York Times", 19 I 2014, SR6. Dane z bardziej dokładnych wyszukiwań można znaleźć za pomocą Google AdWords. Szacunkowe dane można również znaleźć na Google Trends, porównując wyszukiwania zawierające słowa „utalentowany" + „syn" oraz „utalentowana" + „córka". Por. np. https://www.google.com/trends/explore?date=all&geo=US&q=gifted_son,gifted_daughter z https://www.google.com/trends/explore?date=all&geo=US&q=overweight_son,overweight_daughter.

Jedyny wyjątek od wzoru, według którego jest więcej zapytań o intelekt synów i wygląd córek, stanowi większa liczba pytań o „grubego syna" niż „grubą córkę". Prawdopodobnie ma to związek z popularnością kazirodczego porno, o którym była mowa wcześniej. Ok. 20% wyszukiwań ze słowem „gruby" i „syn" zawierała również słowo „porno".

[28] *Gender Equity in Education: A Data Snapshot*, Office for Civil Rights, U.S. Department of Education, VI 2012, http://www2ed..gov/about/offices/list/ocr/docs/gender-equity-in-education.pdf.

[29] Data Resource Center for Child and Adolescent Health, http://www.childhealthdata.org/browse/survey/results?q=2415&g=455&a=3879&r=1.

[30] Seth Stephens-Davidowitz, *The Data of Hate*. Dane na ten temat można pobrać ze strony sethsd.com w zakładce „Stormfront".

[31] Jak wynika z wpisów w Google'u, zainteresowanie Stormfrontem było w październiku 2016 roku zbliżone do tego z października roku 2015, czyli wyraźnie mniejsze niż w czasie pierwszej elekcji Baracka Obamy. W październiku 2008 roku wzrosło niemal o 60 procent w porównaniu z październikiem roku 2007, a w dniu wyboru Obamy na prezydenta liczba wyszukiwań stron Stormfrontu w Internecie zwiększyła się prawie dziesięciokrotnie. Dzień po zwycięstwie Trumpa na ten hejterski portal wchodzono tylko dwa i pół razy częściej, nastąpił zatem mniej więcej taki wzrost wyszukiwań, jak dzień po wyborze George'a W. Busha w 2004 roku – co w znacznym stopniu może być odzwierciedleniem zainteresowania programami informacyjnymi wśród osób nałogowo interesujących się polityką.

[32] Matthew Gentzkow, Jesse M. Shapiro, *Ideological Segregation Online and Offline*, „Quarterly Journal of Economics" 2011, nr 4.

[33] Eytan Bakshy, Solomon Messing, Lada A. Adamic, *Exposure to Ideologically Diverse News and Opinion on Facebook*, „Science" 2015, nr 6239. Autorzy ustalili, że wśród dziewięciu procent aktywnych użytkowników Facebooka, którzy ujawniają swoje poglądy, 23 procent ich znajomych, również otwarcie deklarujących swoje przekonania, ma poglądy przeciwne, a 28,5 procent informacji odbieranych przez nich na Facebooku pochodzi od osób o przeciwstawnych opiniach. Trudno porównać te liczby bezpośrednio z innymi wyliczeniami dotyczącymi podziałów politycznych w sieci, ponieważ uwzględniają one tylko tę niewielką liczbę użytkowników Facebooka, którzy jawnie głoszą swój światopogląd. Przypuszczalnie użytkownicy ci są bardziej aktywni politycznie i kontaktują się z innymi politycznie aktywnymi internautami, wyznającymi tę samą ideologię. Jeżeli to prawda, to rozbieżności wśród wszystkich użytkowników Facebooka muszą być znacznie większe.

[34] Kolejnym powodem sprawiającym, że media społecznościowe są tak zaskakująco różnorodne ideologicznie, jest ich otwarcie na

niezwykle popularne i szeroko rozpowszechniane teksty, i to bez względu na sympatie polityczne. Zob. Solomon Messing, Sean Westwood, *Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online*, „Communication Research", XII 2014.

[35] Zob. Ben Quinn, *Social Network Users Have Twice as Many Friends Online as in Real Life*, „Guardian", 8 V 2011. Artykuł ten omawia przeprowadzone w 2011 roku przez Cystic Fibrosis Trust badania, z których wynika, że przeciętny użytkownik mediów społecznościowych ma stu dwudziestu jeden wirtualnych znajomych online, ale tylko pięćdziesięciu pięciu prawdziwych znajomych offline. Według badań firmy Pew Research z 2014 roku przeciętny użytkownik Facebooka ma więcej niż trzystu znajomych online. Zob. Aaron Smith, *6 New Facts About Facebook*, 3 II 2014, http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/.

[36] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, Lada Adamic, *The Role of Social Networks in Information Diffusion*, w: *Proceedings of the 21st International Conference on World Wide Web*, 2012.

[37] *Study: Child Abuse on Decline in U.S.*, Associated Press, 12 XII 2011.

[38] Zob. Seth Stephens-Davidowitz, *How Googling Unmasks Child Abuse*, „New York Times", 14 VII 2013, SR5, i *idem, Unreported Victims of Economic Downturn*, mimeo, 2013.

[39] *Stopping Child Abuse: It Begins With You*, „The Arizona Republic", March 26, 2016.

[40] Seth Stephens-Davidowitz, *The Return of the D.I.Y. Abortion*, „New York Times", 6 III 2016, SR2. Dane na ten temat i więcej szczegółów można znaleźć na mojej stronie internetowej sethsd.com w zakładce „Self-Induced Abortion".

[41] Alliance for Audited Media, „Consumer Magazines", http://abcas3.auditedmedia.com/ecirc/magtitlesearch.asp.

[42] Wyliczenia autora z 4 X 2016 roku, przeprowadzone za pomocą Facebook Ads Manager.

[43] *List of Most Popular Websites*, Wikipedia. Według danych wyszukiwarki Alexa, która śledzi przeglądane strony, na dzień 4 IX 2016 roku najpopularniejszym portalem pornograficznym był Xvideos, znajdujący się na pięćdziesiątym siódmym miejscu. Według danych z SimilarWeb, również na 4 IX 2016 roku, najpopularniejszym portalem pornograficznym też był Xvideos, uplasował się on jednak na siedemnastej pozycji. Dziesięć najpopularniejszych portali internetowych według Alexa to Google, YouTube, Facebook, Baidu, Yahoo!, Amazon, Wikipedia, TencentQQ, Google India i Twitter.

[44] Podaję tę opowieść za Davidem Kirkpatrickiem; David Kirpatrick, *The Facebook Effect*, New York 2010 (wyd. polskie: *Efekt Facebooka*, przeł. Michał Lipa, Warszawa 2011).

[45] Peter Thiel, Blake Masters, *Zero to One. Notatki o start-upach*, przeł. Konrad Pawłowski, Warszawa 2015.

[46] Rozmawiałem z Xavierem Amatriainem przez telefon 5 V 2015.

[47] Autorska analiza danych z Google Trends.

[48] *The President Speaks at the Islamic Society of Baltimore*, YouTube, 3 II 2016, https://www.youtube.com/watch?v=LRRV-dVqAjdw.

[49] Wg analizy danych z Google Trends, przeprowadzonej przez autora. Ilość zapytań „zabić muzułmanów" była mniejsza niż w podobnym okresie tydzień wcześniej. Dodatkowo mniej było zapytań zawierających słowo „muzułmanie" i jedno z pięciu najbardziej negatywnych słów używanych wobec tej grupy.

ROZDZIAŁ PIĄTY: ZBLIŻENIE

[1] Seth Stephens-Davidowitz, *They Hook You When You're Young*, „New York Times", 20 IV 2014, SR5. Dane i kody tych badań są dostępne na

mojej stronie internetowej, sethsd.com, w dziale „Baseball".

[2] Yair Ghitza i Andrew Gelman, *The Great Society, Reagan's Revolution, and Generations of Presidential Voting*, niepublikowany maszynopis.

[3] Rozmawiałem z Rajem Chettym przez telefon 30 VII 2015.

[4] Raj Chetty et. al., *The Association Between Income and Life Expectancy in the United States, 2001–2014*, JAMA 2016, nr 16.

[5] Julia Belluz, *Income Inequality Is Chipping Away at Americans' Life Expectancy*, vox.com, 11 IV 2016.

[6] Raj Chetty, John Friedman, Emmanuel Saez, *Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings*, „American Economic Review" 2013, nr 7.

[7] Ten fragment rozdziału pochodzi z mojego felietonu *The Geography of Fame*, „New York Times", 23 III 2014, SR6. Dane na temat miejsc urodzenia sławnych ludzi można znaleźć na stronie sethsd.com w zakładce „Wikipedia Birth Rate, by County". Za pomoc w pobraniu całej Wikipedii i zakodowaniu hrabstw, w których urodzili się wszyscy Amerykanie, mający w niej swoje hasła, dziękuję Noahowi Stephensowi-Davidowitzowi.

[8] Więcej dowodów na znaczenie dużych miast w tej kwestii można znaleźć w książce Eda Glaesera, *Triumph of the City*, New York 2011. Glaeser był moim doradcą merytorycznym na studiach doktoranckich.

[9] David Levinson, red., *Encyclopedia of Crime and Punishment* (SAGE, Thousand Oaks, CA, 2002).

[10] Craig Anderson et al., *The Influence of Media Violence on Youth*, „Psychological Science in the Public Interest" 2003, nr 4.

[11] Gordon Dahl i Stefano DellaVigna, *Does Movie Violence Increase Violent Crime*?, „Quarterly Journal of Economics" 2009, nr 2.

[12] Seth Stephens-Davidowitz, *Days of Our Digital Lives*, „New York Times", 5 VII 2015, SR4.

[13] Anna Richardson, Tracey Budd, *Young Adults, Alcohol, Crime and Disorder*, „Criminal Behaviour and Mental Health" 2003, nr 1; Richard A. Scribner, David P. MacKinnon, James H. Dwyer, *The Risk of Assaultive Violence and Alcohol Availability in Los Angeles County*, „American Journal of Public Health" 1995, nr 3; Dennis M. Gorman, Paul W. Speer, Paul J. Gruenewald i Erich W. Labouvie, *Spatial Dynamics of Alcohol Availability, Neighborhood Structure and Violent Crime*, „Journal of Studies on Alcohol" 2001, nr 5; Tony H. Grubesic, William Alex Pridemore, Dominique A. Williams i Loni Philip-Tabb, *Alcohol Outlet Density and Violence: The Role of Risky Retailers and Alcohol-Related Expenditures*, „Alcohol and Alcoholism" 2013, nr 5.

[14] *Ed McCaffrey Knew Christian McCaffrey Would Be Good from the Start* – „The Herd", film na YouTube, post z 3 XII 2015, https://www.youtube.com/watch?v=boHMmp7DpX0.

[15] Badacze dowiedzieli się więcej dzięki temu, że rozłożyli wspomniane dane o przestępczości na mniejsze jednostki czasowe. Przykład? Liczba zgłoszeń przypadków przemocy domowej wzrasta lawinowo, kiedy w jakimś mieście drużyna gospodarzy przegrywa mecz futbolowy, którego była faworytem. Zob. David Card i Gordon B. Dahl, *Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior*, „Quarterly Journal of Economics" 2011, nr 1.

[16] Bill Simmons, *It's Hard to Say Goodbye to David Ortiz*, espn.com, 2 VI 2009, http://www.espn.com/espnmag/story?id=4223584.

[17] Zagadnienie to omawia Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*, New York, 2012 (wyd. polskie: *Sygnał i szum. Sztuka prognozowania w erze technologii*, przeł. Michał Lipa, Gliwice 2014).

[18] Ryan Campbell, *How Will Prince Fielder Age?*, 28 X 2011, http://www.fangraphs.com/blogs/how-will-prince-fielder-age/.

[19] Dane te pozyskałem dzięki uprzejmości Roba McQuowna z Baseball Prospectus.

[20] Rozmawiałem z Isaakiem Kohanem przez telefon 15 VI 2015.

[21] Rozmawiałem z Jamesem Heywoodem przez telefon 17 VIII 2015.

ROZDZIAŁ SZÓSTY: ŚWIAT TO JEDNO WIELKIE LABORATORIUM

[1] Historię tę opisuje między innymi Brian Christian, *The A/B Test: Inside the Technology That's Changing the Rules of Business*, „Wired", 25 IV 2012, http://www.wired.com/2012/04/ff_abtesting/.

[2] Esther Duflo, Rema Hanna, Stephen P. Ryan, *Incentives Work: Getting Teachers to Come to School*, „American Economic Review" 2012, nr 4.

[3] Ian Parker, *The Poverty Lab*, „New Yorker", 17 V 2010.

[4] Christian, *op.cit.*

[5] Douglas Bowman, *Goodbye, Google*, „stopdesign", 20 III 2009, http://stopdesign.com/archive/2009/03/20/goodbye-google.html.

[6] Eytan Bakshy, *Big Experiments: Big Data's Friend for Making Decisions*, 3 IV 2014, https://www.facebook.com/notes/facebook-data-science/big-experiments-big-datas-friend-for-making-decisions/10152160441298859/. Źródła informacji o badaniach farmaceutycznych można znaleźć w tekście *How many clinical trials are started each year?*, „Quora post", https://www.quora.com/How-many-clinical-trials-are-started-each-year.

[7] Rozmawiałem z Danem Sirokerem przez telefon 29 IV 2015.

[8] Dan Siroker, *How Obama Raised $60 Million by Running a Simple Experiment*, Optimizely blog, 29 XI 2010, https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/.

[9] Gazeta „Boston Globe" udostępniła autorowi testy i ich wyniki. Szczegółowe informacje na temat testów w „Globe" można znaleźć w *„The Boston Globe": Discovering and Optimizing a Value Proposition for Content*, Marketing Sherpa Video Archive, https://www.marketingsherpa.com/video/boston-globe-optimizationsummit2. Tekst zawiera też nagranie rozmowy Petera Doucette'a z „Globe" i Pameli Markey z MECLABS.

[10] Rozmawiałem z Clarkiem Bensonem przez telefon 23 VII 2015.

[11] *Enhancing Text Ads on the Google Display Network*, Inside Ad-Sense, 3 XII 2012, https://adsense.googleblog.com/2012/12/enhancing-text-ads-on-google-display.html.

[12] Zob. na przykład „Duże strzałki w reklamach Google'a – prosimy je usunąć", DoubleClick Publisher Help Forum, https://productforums.google.com/forum#!topic/dfp/pTRMqWUF9s.

[13] Adam Alter, *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*, New York 2017 (wyd. polskie: *Uzależnienia 2.0. Dlaczego tak trudno oprzeć się nowym technologiom*, przeł. Aleksander Gomola, Kraków 2018).

[14] Autorska analiza danych z Google Trends.

[15] Problem ten jest omawiany na filmie aktualnie dostępnym na stronie Freakonomics, prowadzonej przez Harry Walker Speakers Bureau, http://www.harrywalker.com/speakers/authors-of-freakonomics/.

[16] Wesley R. Hartmann, Daniel Klapper, *Super Bowl Ads*, niepublikowany maszynopis, 2014.

[17] Aby zapoznać się z przekonującą argumentacją wspierającą tezę, że być może żyjemy w świecie symulacji komputerowej, zob. Nick Bostrom, *Are We Living in a Computer Simulation?*, „Philosophical Quarterly" 2003, nr 211.

[18] Artykuł redakcyjny w „Los Angeles Times", *U.S. Presidential Assasinations and Attempts*, „Los Angeles Times", 22 I 2012, http://timelines.latimes.com/us-presidential-assasinations-and-attempts/.

[19] Benjamin F. Jones, Benjamin A. Olken, *Do Assasins Really Change History?*, „New York Times", 12 IV 2015, SR12.

[20] Wstrząsające nagranie z tego zamachu, zatytułowane *Parade Surprise (Chechnya 2004)*, można zobaczyć na YouTube, post z 31 III 2009, https://www.youtube.com/watch?v=fHWhs5QkfuY.

[21] Benjamin F. Jones i Benjamin A. Olken, *Hit or Miss? The Effect of Asssinations on Institutions and War*, „American Economic Journal: Macroeconomics" 2009, nr 2.

[22] Taką tezę stawia John Tierney, *How to Win the Lottery (Happily)*, „New York Times", 27 V 2014, D5. Autor omawia następujące rozprawy: Bénédicte Apouey, Andrew E. Clark, *Winning Big but Feeling No Better? The Effect of Lottery Prizes on Physical and Mental Health*, „Health Economics" 2015, nr 5; Jonathan Gardner, Andrew J. Oswald, *Money and Mental Wellbeing: A Longitudinal Study of Medium-Sized Lottery Wins*, „Journal of Health Economics" 2007, nr 1; Anna Hedenus, *At the End of the Rainbow: Post-Winning Life Among Swedish Lottery Winners*, niepublikowany maszynopis, 2011. Tierney zauważa też, że słynny artykuł z 1978 roku (Philip Brickman, Dan Coates, Ronnie Janoff-Bulman, *Lottery Winners and Accident Victims: Is Happiness Relative?*, „Journal of Personality and Social Psychology" 1978, nr 8), dowodzący, że wygrana na loterii w ogóle nie zapewnia szczęścia zwycięzcom, opierał się na badaniach wyjątkowo nielicznej grupy osób.

[23] Zob. Peter Kuhn, Peter Kooreman, Adrian Soetevent, Arie Kapteyn, *The Effects of Lottery Prizes on Winners and Their Neighbors: Evidence from the Dutch Postcode Lottery*, „American Economic Review" 2011, nr 5, oraz Sumit Agarwal, Vyacheslaw Mikhed i Barry Scholnick, *Does Inequality Cause Financial Distress? Evidence from Lottery Winners and Neighboring Bankruptcies*, artykuł roboczy, 2016.

[24] Agarwal, Mikhed, Scholnick, *op.cit.*

[25] Jeffrey Clemens i Joshua D. Gottlieb, *Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?*, „American Economic Review" 2014, nr 4. Proszę zauważyć, że z tego wniosku wcale nie wynika, jakoby lekarze byli po prostu niegodziwcami. Prawdę mówiąc, należałoby ich oceniać gorzej, gdyby dodatkowe zabiegi i badania, zalecane przez lepiej opłacanych medyków, rzeczywiście przyczyniały się znacząco do ratowania życia pacjentów – bo gdyby tak było, oznaczałoby to, że lekarzom należy płacić więcej, aby ordynowali pacjentom zabiegi i badania ratujące życie. Wyniki analizy Clemensa i Gottlieba wskazują jednak, że lekarze zalecają takie – to znaczy ratujące życie – badania bez względu na to, ile otrzymują na nie pieniędzy. Trzeba jednak ich opłacać sowiciej, jeśli chcemy, aby kierowali podopiecznych na zabiegi i badania, które wcale im zbytnio nie pomagają. Inaczej mówiąc, w sytuacji zagrożenia życia pacjenta lekarze raczej nie kierują się motywacją finansową, ma ona jednak dla nich bardzo duże znaczenie w wypadku badań czy zabiegów, które nie mogą szczególnie pomóc choremu.

[26] Robert D. McFadden, Eben Shapiro, *Finally, a Face to Fit Stuyvesant; a High School of High Achievers Gets a High-Priced Home*, „New York Times", 8 IX 1992.

[27] Oferta programowa szkoły znajduje się na stronie internetowej Stuya, http://stuy.enschool.org/index.jsp.

[28] Anna Bahr, *When the College Admissions Battle Starts at Age 3*, „New York Times", 29 VII 2014, http://www.nytimes.com/2014/07/30/upshot/when-thecollege-admissions-battle-starts-at-age-3.html.

[29] Sewell Chan, *The Obama Team's New York Ties*, „New York Times", 25 XI 2008; Evan T.R. Rosenman, *Class of 1984: Lisa Randall*, „Harvard Crimson", 2 VI 2009; *Gary Shteyngart on Stuyvesant High School: My New York*, YouTube, 4 VI 2010, https://www.youtube.com/watch?v=NQ_phGkC-Tk; Candace Amos, *30 Stars Who Attended NYC Public Schools*, „New York Daily News", 29 V 2015.

[30] Carl Campanile, *Kids Stuy High Over Bubba; He'll Address Ground Zero School's Graduation*, „New York Post", 22 III 2002; United Nations Press Release, *Stuyvesant High School's „Multicultural Tapestry" Eloquent Response to Hatred, Says Secretary-General in Graduation Address*, 23 VI 2004; *Conan O'Brien's Speech at Stuyvesant's Class of 2006 Graduation in Lincoln Center*, YouTube, 6 V 2012, https://www.youtube.com/watch?v=zAMkUE9Oxnc.

[31] Zob. https://k12.niche.com/rankings/public-high-schools/best-overall/.

[32] Pamela Wheaton, *8th-Graders Get High School Admissions Results*, „Insideschools", 4 III 2016, http://insideschools.org/blog/item/1001064-8th-graders-get-high-school-admission-results.

[33] M. Keith Chen, Jesse M. Shapiro, *Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach*, „American Law and Economics Review" 2007, nr 1.

[34] Atila Abdulkadiroğlu, Joshua Angrist i Parag Pathak, *The Elite Illusion: Achievement Effects at Boston and New York Exam Schools*, „Econometrica" 2014, nr 1. Do identycznych wniosków, czyli że nauka w prestiżowej szkole właściwie nic nie daje, doszli niezależnie od tych autorów Will Dobbie i Roland G. Fryer Jr., *The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools*, „American Economic Journal: Applied Economics" 2014, nr 3.

[35] Zob. http://www.payscale.com/college-salary-report/bachelors.

[36] Stacy Berg Dale, Alan B. Krueger, *Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables*, „Quarterly Journal of Economics" 2002, nr 4.

[37] Alice Schroeder, *The Snowball: Warren Buffet and the Business of Life*, New York 2008.

ROZDZIAŁ SIÓDMY: BIG DATA, BIG SRATA? CZEGO NIE POTRAFI ZBIÓR BIG DATA

[1] Johan Bollen, Huina Mao, Xiaojun Zeng, *Twitter Mood Predicts the Stock Market*, „Journal of Computational Science" 2011, nr 1.

[2] James Mackintosh, *Hedge Fund That Traded Based on Social Media Signals Didn't Work Out*, „Financial Times", 25 V 2012.

[3] Christopher F. Chabris et al., *Most Reported Genetic Associations with General Intelligence Are Probably False*, „Psychological Science" (2012).

[4] Przypadek Zoë Chance został przedstawiony w TEDx Talks, *How to Make a Behavior Addictive: Zoë Chance at TEDx Mill River*, YouTube, 14 V 2013, https://www.youtube.com/watch?v=AHfiKav9fcQ. Niektóre szczegóły, jak kolor pedometru, zostały zaakcentowane w filmie. Rozmawiałem z Zoë Chance przez telefon 20 IV 2015 i przeprowadziłem z nią wywiad e-mailowy 11 VII oraz 8 IX 2016.

[5] Ten fragment pochodzi z artykułu Alexa Peysakhovicha i Setha Stephensa-Davidowitza, *How Not to Drown in Numbers*, „New York Times", 3 V 2015, SR6.

[6] Brian A. Jacob i Steven D. Levitt, *Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating*, „Quarterly Journal of Economics" 2003, nr 3.

[7] Rozmawiałem z Thomasem Kanem przez telefon 22 IV 2015 roku.

[8] Bill & Melinda Gates Foundation, *Ensuring Fair and Reliable Measures of Effective Teaching*, http://k12education.gatesfoundation.org/wp-content/uploads/2015/05/MET_Ensuring_Fair_andReliable-Measures_Practitioner_Brief.pdf.

ROZDZIAŁ ÓSMY: WIĘCEJ DANYCH, WIĘCEJ PROBLEMÓW? CZEGO NIE POWINNIŚMY ROBIĆ

[1] Oded Netzer, Alain Lemaire, Michal Herzenstein, *When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications*, 2016.

[2] Peter Renton, *Another Analysis of Default Rates at* Lending Club *and* Prosper, 25 X 2012, http://www.lendacademy.com/lending-club-prosper-default-rates/.

[3] Michal Kosinski, David Stillwell, Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, „PNAS" 2013, nr 15.

[4] Michael Luca, *Reviews, Reputation and Revenue: The Case of Yelp*, niepublikowany maszynopis, 2011.

[5] Christine Ma-Kellams, Flora Or, Ji Hyun Baek, Ichiro Kawachi, *Rethinking Suicide Surveillance: Google Search Data and Self Reported Suicidality Differentially Estimate Completed Suicide Risk*, „Clinical Psychological Science" 2016, nr 3.

[6] Zastosowałem tu taką samą metodę jak ta, którą przedstawiłem na swoim portalu w komentarzach do samodzielnie przeprowadzanej aborcji, a wyszukiwania dotyczące samobójstwa zestawiłem z wyszukiwaniami dotyczącymi sposobów wiązania krawata. W 2015 roku pytania typu „Jak się wiąże krawat?" pojawiły się w Google'u sześć milionów sześćset tysięcy razy, wyszukiwań w kategorii „samobójstwo" odnotowano natomiast sześć i pół razy więcej.

[7] Bridge Initiative Team, *When Islamophobia Turns Violent: The 2016 U.S. Presidential Election*, 2 V 2016, tekst dostępny na http://bridge.georgetown.edu/when-islamophobia-turns-violent-the2016-u-s-presidential-elections/.

ZAKOŃCZENIE: ILU LUDZI CZYTA KSIĄŻKI DO KOŃCA?

[1] Karl Popper, *Conjectures and Refutations*, London 1963 (wyd. polskie: *Droga do wiedzy: Domysły i refutacje*, przeł. Stefan Amsterdamski, Warszawa 1999).

[2] Simon Rogers, *John Snow's Data Journalism: The Cholera Map That Changed the World*, „Guardian", 15 III 2013.

[3] Rozmawiałem z Benjaminem Jonesem przez telefon 1 VI 2015. Jego badania opisuje również artykuł Aarona Chatterjiego i Benjamina Jonesa, *Harnessing Technology to Improve K-12 Education*, „Hamilton Project Discussion Paper", 2012.

[4] Jordan Ellenberg, *The Summer's Most Unread Book Is...*, „Wall Street Journal", 3 VII 2014.

[1*] Wyszukiwarka Google Trends stała się źródłem znacznej części danych do moich analiz. Mimo to, ponieważ pozwala jedynie porównywać relatywną częstotliwość wyszukiwania i nie podaje liczb absolutnych na temat żadnego pojedynczego hasła, zwykle uzupełniałem te informacje za pomocą Google AdWords, która podaje dokładnie, jak często w sieci pojawiają się konkretne wpisy. W większości wypadków byłem w stanie wyostrzyć ten informacyjny obraz jeszcze bardziej, pomagając sobie opracowanym samodzielnie algorytmem, opartym na Google Trends, który przedstawiam w mojej rozprawie doktorskiej, *Essays Using Google Data*, i w artykule opublikowanym w „Journal of Public Economics" – *The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data*. Rozprawa, link do artykułu i pełne wyjaśnienie danych oraz kodu, jaki wykorzystywałem w badaniach przedstawionych w niniejszej książce, dostępne są na mojej stronie internetowej: sethsd.com.

[2*] Malcolm Gladwell, *Błysk! Potęga przeczucia*, przeł. Anna Skucińska, Kraków 2019 (przyp. tłum.).

[3*] Odkrywam się przed wami całkowicie: niedługo po tym, jak przeprowadziłem to badanie, przeniosłem się z Kalifornii do Nowego Jorku. Używanie danych, żeby się dowiedzieć, co należy robić, bywa łatwe – trudniej jest zastosować się do własnych rad.

[4*] Wprawdzie pierwsza wersja Google Flu miała poważne wady, ale informatycy niedawno ulepszyli tę wyszukiwarkę i dziś jest ona sprawniejsza.

[5*] Gdy w 1998 roku ktoś wpisywał hasło „samochody" w wyszukiwarkę starszą niż Google, to zalewały go adresy portali pornograficznych[7].

Portale te często miały wpisane na swojej stronie startowej słowo „samochody” białymi literami na białym tle, by oszukiwać wyszukiwarki internetowe, i dzięki temu na strony erotyczne wchodzili czasem ludzie, którzy wprawdzie chcieli kupić nowy samochód, ale których uwagę przypadkowo zajęła też na chwilę pornografia.

[6*] STD (*sexually transmitted disease*) – choroba weneryczna (przyp. tłum.).

[7*] Oto kolejna teoria, nad którą pracuję: big data potwierdza wszystko, co mówił nieżyjący już Leonard Cohen, który między innymi poradził kiedyś siostrzeńcowi, jak podrywać kobiety: „Słuchaj uważnie. Potem jeszcze więcej. A kiedy pomyślisz, że masz już tego dość, to słuchaj jej dalej”[21]. Wydaje się, że wyniki badań naukowych przynajmniej z grubsza potwierdzają słuszność tezy artysty.

[8*] Jeszcze jeden powód, dla którego kłamiemy w sondażach, jest taki, że chcemy po prostu trochę narozrabiać, powygłupiać się[6]. Zazwyczaj nie jest to poważny problem, wyjąwszy jednak sondaże wśród nastolatków, co mocno utrudnia nam zrozumienie postaw tej grupy wiekowej. Na przykład kiedyś sądzono, że młodzi ludzie, którzy wychowywali się w rodzinach adopcyjnych, mają większe skłonności do nagannych zachowań, takich jak zażywanie narkotyków, spożywanie alkoholu czy wagary. Później jednak okazało się, że można całkowicie wyjaśnić tę korelację, biorąc pod uwagę, że dziewiętnaście procent rzekomo adoptowanych młodych ludzi wcale nie pochodziło z rodzin adopcyjnych. Dalsze badania wskazują, że znaczny odsetek nastolatków odpowiada w ankietach, że mierzą więcej niż siedem stóp wzrostu, ważą ponad czterysta funtów albo są rodzicami trojga dzieci. Przeprowadzono też badanie, z którego wynika, że dziewięćdziesiąt dziewięć procent nieletnich, którzy podali w sondażu, że mają sztuczną kończynę, po prostu żartowało.

[9*] Być może niektórzy czytelnicy uznają za obraźliwe, że nawet żartem kojarzę męskie upodobanie do Judy Garland z pragnieniem uprawiania seksu z mężczyznami, chociaż wcale nie chcę tego sugerować. I na pewno nie chcę też sugerować, że wszyscy geje czy

chociażby ich większość fascynuje się diwami. Dane analityczne wskazują jednak, że w tym stereotypie tkwi ziarno prawdy. Z moich badań wynika, że mężczyźni, którzy szukają w sieci informacji na temat Judy Garland, trzykrotnie częściej wyszukują też pornografię gejowską niż heteroseksualną. Jak zatem dowodzi big data, niektóre stereotypy bywają uzasadnione.

[10*] Wydaje mi się, że dane te mają też znaczenie dla przyjęcia optymalnej strategii randkowej, wyraźnie bowiem wychodzi przecież na to, że powinniśmy się jak najczęściej umawiać, często dostawać kosza i za bardzo się tym nie przejmować, ponieważ prędzej czy później każdy z nas znajdzie sobie partnerkę czy partnera, zainteresowanego dokładnie kimś takim jak my. Bez względu na nasz wygląd są gdzieś na świecie ludzie, którym się spodobamy. Wierzcie mi, tak jest.

[11*] Początkowo chciałem nazwać tę książkę *Jak dużego mam członka, czyli co wyszukiwania w Google'u mówią o naturze ludzkiej*, wydawca ostrzegł mnie jednak, że pod takim tytułem będzie ją trudno sprzedawać, bo czytelnicy mogliby się za bardzo krępować, by kupić takie dzieło na przykład w księgarni na lotnisku. Zgadzacie się z tą opinią?

[12*] Aby lepiej zweryfikować hipotezę, że rodzice traktują dzieci różnie ze względu na ich płeć, próbuję pozyskać też dane z portali rodzicielskich, gdzie dotarłbym do znacznie większej grupy rodziców niż jedynie do tych, którzy dokonują takich konkretnych, sprecyzowanych wpisów.

[13*] Analizowałem dane z Twittera. Dziękuję Emmie Pierson za pomoc w ich ściągnięciu. Nie brałem pod uwagę – jako mało odkrywczych – doniesień kobiet o tym, co ich mężowie robią w danej chwili, chociaż takich postów w mediach społecznościowych jest najwięcej. Te posty też jednak brzmią raczej przychylnie dla mężczyzn, bo panie donoszą najczęściej, że mąż akurat „pracuje" albo „gotuje".

[14*] W tym miejscu odsłaniam się przed wami całkowicie: kiedy przeprowadzałem faktograficzną korektę tej książki, Noah zaprzeczył, jakoby niechęć do ulubionego sportu Amerykanów stanowiła podstawę

jego osobowości. Przyznaje, że go nie cierpi, ale uważa, że jego tożsamość charakteryzują przede wszystkim dobroć, miłość do dzieci i inteligencja, stosunek do bejsbolu zaś nie zmieściłby się nawet w pierwszej dziesiątce. Uznałem jednak, że czasem trudno przychodzi nam spojrzeć na siebie obiektywnie, a ja jako obserwator zewnętrzny widzę, że wrogość wobec bejsbolu jest jednak fundamentalną cechą osobowości Noaha bez względu na to, czy on sam to dostrzega, czy też nie. Dlatego nie usunąłem tego fragmentu tekstu.

[15*] Mam przykład, który pokazuje, jak coś, co na pierwszy rzut oka wydaje się złem, da się uznać za dobro, jeśli zapobiega czemuś gorszemu. Ed McCaffrey, absolwent Uniwersytetu Stanforda i były futbolista, używa następujących argumentów, aby uzasadnić, dlaczego wszystkim swoim czterem synom pozwala grać w futbol[14]: „Moi chłopcy są bardzo żywi. Więc jeśli nie grają w futbol, to jeżdżą na skejtbordach, chodzą po drzewach, grają w berka w ogrodzie albo bawią się w paintball. Rozumiesz, nie będą po prostu siedzieć i nic nie robić. Więc ja patrzę na to tak, że hej, w futbolu obowiązują przynajmniej jakieś zasady… Moi synowie lądowali w szpitalu, bo spadali z tarasów, miewali wypadki rowerowe czy skejtbordowe albo spadali z drzew. Wiesz, robili różne rzeczy… Owszem, futbol to pełen przemocy, kontaktowy sport. Ale moi chłopcy mają do niego odpowiednią osobowość, a przynajmniej nie skaczą po górach w wingsuitach i nie popełniają innych podobnych wariactw. W futbolu mówimy, jak mi się wydaje, o kontrolowanej agresji”. Nigdy wcześniej nie słyszałem podobnej argumentacji. McCaffrey posłużył się nią w programie *The Herd with Colin Cowherd*, ale po lekturze raportu Dahla/DellaVigni traktuję ją poważnie. Zaletą wielkich zbiorów danych ze świata realnego, a nie z laboratoriów, jest to, że można dzięki nim dostrzec niewidoczne inaczej konsekwencje.

[16*] Jeśli dotarliście do tej części książki, możecie już pewnie zgadnąć, że do wszelkich tego rodzaju opowieści o pozytywnej wymowie podchodzę cynicznie, chciałem umieścić tu jednak chociażby jedną taką krzepiącą anegdotę, więc ograniczam swój cynizm do przypisu. Otóż wydaje mi się, że dzięki metodzie PECOTA udało się ustalić tylko tyle, że

Ortiz najpierw używał sterydów, potem przestał, a później do nich wrócił. Z punktu widzenia prognostyki jest chyba całkiem cool, że PECOTA dała taki wynik – ale historia Ortiza staje się jak gdyby mniej wzruszająca.

[17*] Słynny artykuł z 1978 roku, głoszący, że wygrana na loterii w ogóle nie daje szczęścia, został w znacznej mierze zdezawuowany.

[18*] Zmieniłem nazwisko mojego bohatera i kilka szczegółów.

[19*] Milan Kundera, *Nieznośna lekkość bytu*, przeł. Agnieszka Holland, Warszawa 1985 (przyp. tłum.).

[20*] Szukając osób takich jak Yilmaz, którym niewiele brakowało, żeby zdać egzamin, zdumiałem się niepomiernie, jak wielu ludzi – mniej więcej pomiędzy dwudziestym i pięćdziesiątym rokiem życia – dobrze pamięta ten egzamin wstępny z młodych lat i nadal opowiada o swoim niepowodzeniu jak o dramacie życiowym. Wśród takich moich rozmówców znalazł się na przykład były kongresmen i kandydat na burmistrza Nowego Jorku Anthony Weiner, któremu, jak sam twierdzi, do zdania egzaminu zabrakło jednego punktu. „Nie chcieli mnie w tej szkole", powiedział mi w rozmowie telefonicznej.

[21*] Steven D. Levitt, Stephen J. Dubner, *Freakonomia. Świat od podszewki*, przeł. Agnieszka Sobolewska, Kraków 2011 (przyp. tłum.).

[22*] Friedrich Nietzsche, *Poza dobrem i złem*, przeł. Stanisław Wyrzykowski, Kraków 2010 (przyp. tłum.).

[23*] Ponieważ wszyscy kłamią, powinniście zakwestionować większą część tej historii. Może wcale nie jestem taki wytrwały. Może nie pracowałem nad tą książką tak ciężko. Może podobnie jak wielu ludzi przesadzam, opowiadając wam, jak dużo pracuję. Może wśród moich trzynastu miesięcy „ciężkiej pracy" było kilka, kiedy wcale nie pisałem. Może nie żyłem jak pustelnik. Może gdybyście sprawdzili mój profil na Facebooku, zobaczylibyście mnie na zdjęciach z tamtego czasu w towarzystwie przyjaciół, chociaż rzekomo żyłem wtedy samotnie. A może owszem, żyłem samotnie, ale nie z własnej woli. Może

spędziłem samotnie niejedną noc, nie mogąc pracować i mając płonną nadzieję, że ktoś się do mnie odezwie. Może nikt nie przysyła mi żadnych zaproszeń. Może nikt nie pisze do mnie na Bumble'u. Wszyscy kłamią. Narratorzy godni zaufania nie istnieją.

[24*] Donna Tartt, *Szczygieł*, przeł. Jerzy Kozłowski, Kraków 2015 (przyp. tłum.).

[25*] Thomas Piketty, *Kapitał w XXI wieku*, przeł. Andrzej Bilik, Warszawa 2015 (przyp. tłum.).

[26*] Carl Shapiro, Hal. R. Varian, *Potęga informacji. Strategiczny przewodnik po gospodarce sieciowej*, przeł. Katarzyna Masłowska, Cezar Matkowski, Agnieszka Sobolewska, Gliwice 2007 (przyp. tłum.).